打★的是一定要能默写的，没打★的也一定要熟悉！

问题是学长整理的，答案是我自己整理的。

CH1 导论

1.1 什么是统计学?

统计学是一门研究数据的科学，任务是

- 如何有效地收集、整理、分析和解释这些数据;
- 探索数据内在的数量规律性;
- 对所观察的现象做出推断或预测，直到为采取决策提供依据。

2015年重大432真题：统计学是一门关于研究客观事物数量方面与数量关系的（D）

A. 理论统计与运用统计　　B. 统计预测与决策

C. 描述统计与推断统计　　D. 统计资料收集与分析

1.2 数据分析的方法有哪些?

- **描述统计方法**：研究数据**收集、处理、汇总、图表描述、概括与分析等**统计方法。
- **推断统计方法**：研究如何利用样本数据来**推断总体特征**的统计方法。例如，要了解一个地区全部人口特征，不可能对每个人的特征一一测量，需要抽取部分个体——即样本进行测量，然后根据获得的样本数据对所研究的总体特征进行推断。

2021年重大432真题：最近发表的一份报告称：由150部新车价格组成的一份样本表明，外国新车的价格明显高于本国生产的新车。这一结论属于（C）

A. 对样本的描述统计　　B. 对样本的推断统计

C. 对总体的推断统计　　D. 对总体的描述统计

1.3 统计学的应用领域有哪些?

- 企业发展战略
- 产品质量管理
- 市场研究
- 财务分析
- 经济预测
- 人力资源管理

1.4 数据分析的真正目的是?

有些人的心目中可能有了某种结论或希望看到符合他们需要的结论，然后去找这些统计数据来支持他们的结论，这是错误的。

数据分析的真正目的是从数据中找出规律，从数据中寻找启发，而不是寻找支持。真正的数据分析事先是没有结论的，通过对数据的分析才能得出结论。

1.5 ★★★ 统计数据有哪些类型？（按照三种分类方式划分）

- 按被描述的对象与**时间**的关系，可以分为截面数据和时间序列数据。
- 按统计数据的**收集方法**，可以分为观测数据和实验数据。
- 按所采用的**计量尺度**的不同，可以分为：
    1. (无序) 分类数据，也称定性数据，是由分类尺度计量形成的。分类数据是只能归于某一类别的非数字型数据，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，例如“上市公司所属行业”和“商品的产地”。
    2. 顺序数据，是“有序的分类型数据”，例如“成绩的层次”和“受教育的程度”。和分类数据统称品质数据。
    3. 数值型数据，也称为定量数据，是使用自然或度量衡单位对事物进行测量的结果，数据表现为具体的数值，例如“气温”。

1.6 为什么要区分不同的数据类型?

因为对不同类型的数据，我们会采用不同的统计方法来处理和分析。

例如，对分类数据可以计算出各组的频数或频率，计算其众数与异众比率，进行列联表的χ^2分析等。

对顺序数据，可以计算其中位数和四分位差，计算等级相关系数等。

对数值型数据，可以用更多的统计方法进行处理，例如计算各种统计量、进行参数的估计与检验等。

1.7 简述总体和样本的含义（重点理解总体的含义）

- 总体是包含所研究的全部个体（数据）的集合，是我们所关心的一些个体组成。例如多个人构成的集合就是总体，这其中的一个人就是个体。

  2014年重大432真题参考答案：统计上的总体通常是一组观测数据，而不是一群人或一些物品的集合（数据的科学）。
- 样本是从总体中抽取的一部分元素的集合。

1.8 简述参数和统计量的含义

- 参数是用来描述总体特征的概括性数字度量。
- 统计量是用来描述样本特征的概括性数字度量。统计量是样本的函数，与未知参数无关。

1.9 简述变量及其含义

变量是说明现象某种特征的概念，特点是从一次观察到下一次观察会呈现出差别或变化。

Ch2 数据的搜集

2.1 数据的来源有哪些?

- 直接来源，直接来源主要有两个渠道：一是调查或观察，二是实验。
- 间接来源，主要是公开出版或公开报道的数据。

2.2 二手数据的优缺点有哪些?

- 优点：搜集方便，数据采集快，成本低，而且有些资料还可以提供研究问题和背景，帮助研究问题。
- 缺点：针对性不够好，例如可能口径不一致，数据过时。

2.3 ★★★ 如何评估二手数据?

3W1H原则

who	why	how	when
谁收集的	出于什么目的收集	怎样收集的	何时收集的

2.4 简述普查的定义

普查是指为特定目的而专门组织的全面调查，例如工业普查、农业普查和人口普查等。

2.5 如何评判一个好的样本?

- 针对研究目的，一个好的样本应贴合研究目的。
- 保证数据质量的前提下，尽量降低调查成本。

2.6 ★★★ 简述概率抽样和非概率抽样的特点和区别

（需要先说明定义，再展开说明其包含的抽样方式以及每个抽样方式的定义和优劣）

概率抽样也称为随机抽样，是指遵循随机原则进行的抽样，总体中每个单位都有一定机会入样。概率抽样有以下几种特点：

1. 抽样时按一定的概率以**随机原则**抽取样本，排除主观上的有意识抽样，**使每个单位都有一定的机会被选中。**
2. 每个单位被抽中的**概率是已知的或可计算的。**
3. 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。**

概率抽样包含的抽样方式：

1 简单随机抽样：是指从包括 N 个总体单位的抽样框中独立随机的，一个一个的抽取样本量为 n 的样本，每个单位被抽中的概率均相同。简单随机抽样是其他抽样方法的基础，其优点是简单直观，当抽样框完整时，可以根据样本统计量直接对总体参数进行估计。缺点是当 n 很大时候，构造这样的抽样框并不容易。

2 分层抽样：是指将抽样单位按照某种标志或规则划分为不同的层，然后从不同的层中独立随机的抽取样本，使得样本中含有各种特征的单位，这种做法使得样本中的结构与总体中的结构比较接近，因而提高了估计的精度，另外分层抽样在某些条件下，实施起来比较方便，但多适用于层间差异大而层内差异小的总体。

3 整群抽样：将总体若干个单位合并成组，这样的组成为群。抽样时直接抽取群中所有个体，然后对其实施调查。整群抽样优点是，抽样时只需群的抽样框，简化了工作量。调查地点相对比较集中，因而节约了调查费用。但整群抽样的显著缺点是，在样本量 n 一定的条件下，整群抽样的估计精度较差。

4 系统抽样：系统抽样是指将总体中所有抽样单位按照一定的顺序排列好，从中随机抽取出一个初始抽样单位，然后按照规则继续依此确定其他抽样单位。典型的例子是从数字 1 到 k 中随机抽取数字 r，然后依此 r+k，…优点操作简便，若有其他辅助信息，对总体内单位有组织的排列。则可以有效提高估计的精度，在实际调查中有广泛应用。缺点是对估计量的方差估计比较困难。

5 多阶段抽样：类似于整群抽样，先从总体中抽取若干群，不再对群中所有单位进行调查，而是进一步抽样进行调查，此时成为二阶段抽样。如果将方法推广，使得抽样的段数增多，则成为多阶段抽样。多阶段抽样的优点是：保证了样本相对比较集中，节约了调查费用。同时特点是不必调查所有初级抽样单位，由于实行了在抽样，可在较大范围调查中应用广泛。

补：分层抽样与整群在抽样的误差来源

- 分层抽样应使组间差异尽可能大、组内差异尽可能小，误差来源主要是组间误差。
- 整群抽样应要求各群有良好的代表性，即组内差异尽可能大、组间差异尽可能小，误差来源主要是群内误差。

非概率抽样是指抽样时没有依据随机原则，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行调查。

非概率抽样包含的抽样方式：

6 方便抽样：是指调查员依据方便的原则，从总体中抽取部分单位对其实施调查。例如，在公共场所等特定空间，如街头等对过往行人进行拦截式的调查，这种抽样方法的显著优点是操作简单且成本低，调查结果可以用于对研究的问题产生一些初步的了解和认识，便于后续更深入的分析。但是，由于样本不具有随机性，因而无法将样本结果推广到总体。

7 判断抽样：是指研究人员根据经验判断和对研究对象的了解，有目的的选取一些样本对其

实施调查。根据研究目的不同，可以分为重点抽样，典型抽样，代表抽样。同时，没有依据随机性原则抽样，样本不具有随机性，因而无法用样本观测结果推断总体。

8 重点抽样：是指从总体中抽取少数重点单位对其实施调查，样本虽然少，但是在总体中却有重要地位。

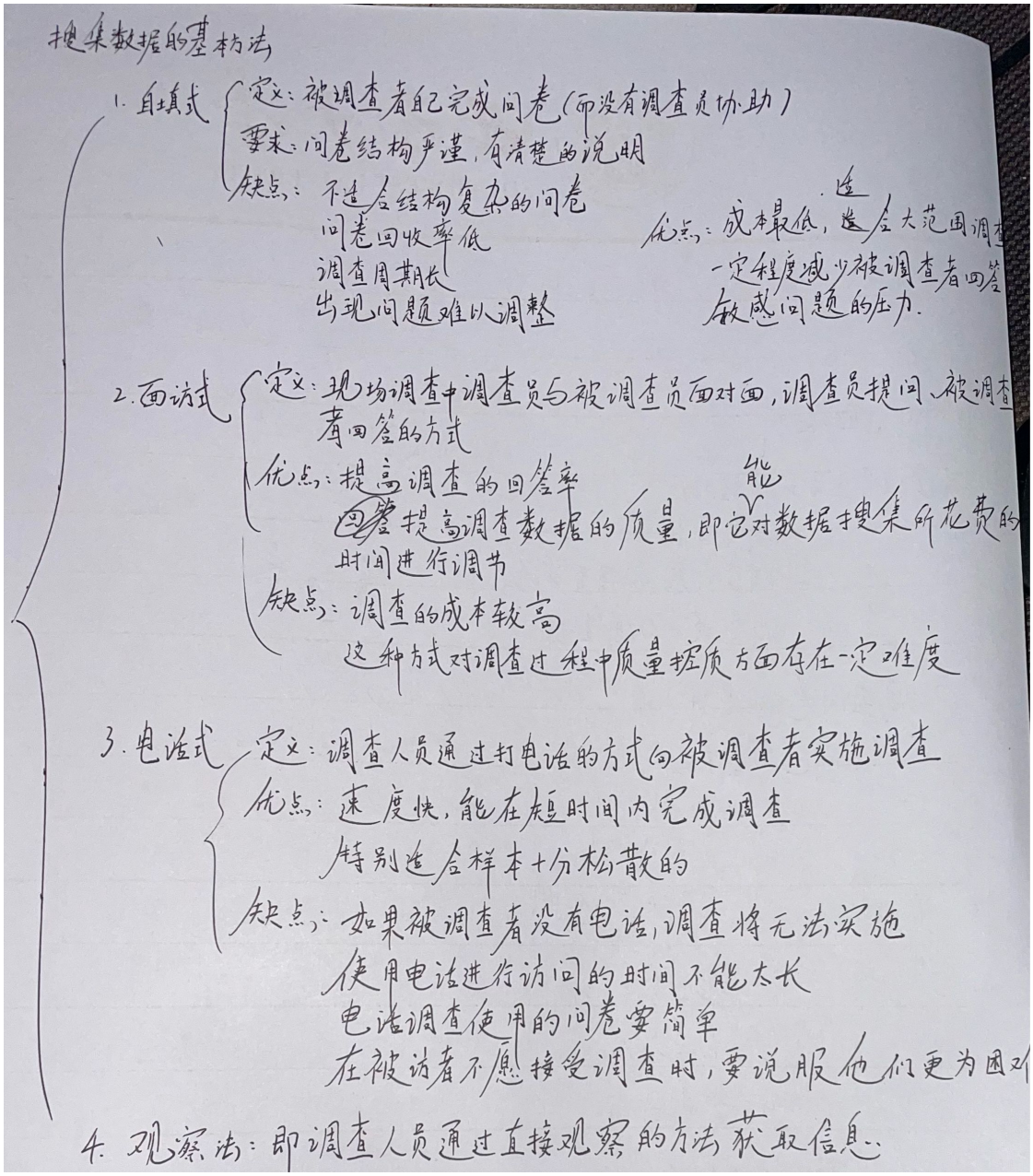
9 典型抽样：是指从总体中抽取部分具有代表性的单位对其进行深入的调研，目的是通过典型单位来描述或揭示所研究问题的本质和规律。因此，所选择的典型单位必须具有研究问题的本质或特征。

10 滚雪球抽样：常用于对稀少群体的调查，在滚雪球抽样中常先选择一组调查单位，对其实施调查以后，在请他们提供另外一些属于研究总体的调查对象，然后实行调查，过程持续下去则形成滚雪球效应。优点是是容易找到属于特定群体的被调查者，但是缺点是由于抽样过程没有依据随机性原则，样本不具有随机性，因此无法使用样本结果推断总体。

11 自愿样本：是指被调查者自愿接受调查成为样本中的一份子，向研究者提供有关信息。例如网络上的许多问卷调查活动。自愿样本由于起不具有随机性，无法用调查的样本结果来推断总体，但其反映了某类群体的一般看法，仍可为研究人员提供大量有用的信息。

12 配额抽样：类似于分层抽样，先将总体按照不同的标志划分为不同的层，然后对各个层采用判断抽样或方便抽样的方法选择一些单位对其实施调查。其优点是，该方法使得样本中包含总体不同类别的样本，即使得样本结构与总体结构比较接近，但是由于抽样不具有随机性，因而无法使用调查的样本信息去推断总体。

2.7 ★★★ 搜集数据的基本方法有哪些?



2.8 选择搜集数据的方法时，应该注意什么问题?

4. 数据搜集方法的选择

搜集数据的不同方法各有特点，在选择数据搜集方法时，需要考虑以下几个问题。

(1) 抽样框中的有关信息。

抽样框中的有关信息是影响方法选择的一个因素。如果抽样框中没有通信地址，就不能将自填式问卷寄给被调查者；如果没有计算机随机数字拨号系统，又没有电话号码的抽样框，电话调查的样本就难以产生，电话访问就无法使用。

(2) 目标总体的特征。

目标总体的特征也会影响数据搜集方法。目标总体的特征表现在多个方面。例如，如果总体的识字率很低，对问卷的理解有困难，就不宜使用自填式方法。样本的地理分布也很重要，如果样本单位分布很广，地域跨度大，进行面访调查的交通费用就会很高，调查过程的管理和质量监控实施起来也不容易。

(3) 调查问题的内容。

调查问题的内容也会影响数据搜集方法。面访调查比较适合复杂的问题，因为调查员可以在现场对模糊的问题进行解释和澄清，并判断被访者对问题是否真正理解，调查问卷的设计也可以采用更多技术，如跳答、转答等，使搜集的数据满足研究的要求。如果调查涉及敏感问题，那么使用匿名的数据搜集方法如自填式或电话式可能更合适。

(4) 有形辅助物的使用。

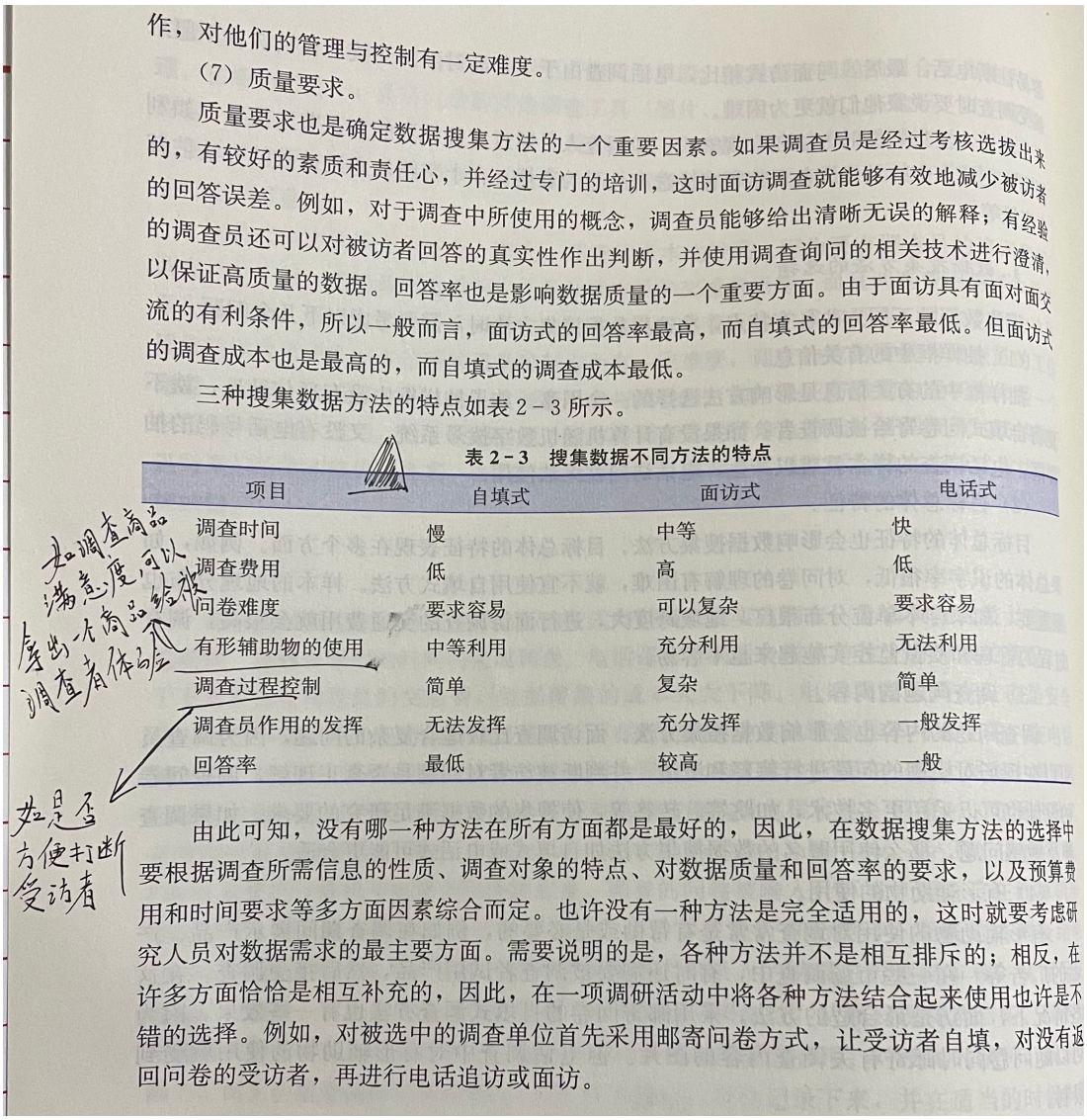
有形辅助物的使用对调查常常是有帮助或是必要的，例如在调查期间展示产品、产品的广告等，在一些市场调查中，有时还需要被调查者试用产品，然后接受调查。在这些情况下，面访是最合适的方法。采用邮寄问卷的自填式调查方法也有一些效果，因为可以随问卷同时邮寄有关调查内容的图片。但电话调查中对有形辅助物的使用就受到限制。

(5) 实施调查的资源。

实施调查的资源会对搜集数据的方法产生重大影响。这些资源包括经费预算、人员、调查设备和调查所需时间。面访调查的费用是最高的，需要支付调查员的劳务费、调查交通费、被访者的礼品费等，还要找到能够满足调查需要的一定数量的调查员。如果使用计算机辅助电话调查，就需要有计算机设备和 CATI 操作系统。

(6) 管理与控制。

有些数据搜集方法比另一些方法更容易管理。例如，在电话调查中，调查员通常集中在调查中心一起工作，因此，管理和控制相对简单。而面访调查中调查员分散、独立地工



作，对他们的管理与控制有一定难度。

（7）质量要求。

质量要求也是确定数据搜集方法的一个重要因素。如果调查员是经过考核选拔出来的，有较好的素质和责任心，并经过专门的培训，这时面访调查就能够有效地减少被访者的回答误差。例如，对于调查中所使用的概念，调查员能够给出清晰无误的解释；有经验的调查员还可以对被访者回答的真实性作出判断，并使用调查询问的相关技术进行澄清，以保证高质量的数据。回答率也是影响数据质量的一个重要方面。由于面访具有面对面交流的有利条件，所以一般而言，面访式的回答率最高，而自填式的回答率最低。但面访式的调查成本也是最高的，而自填式的调查成本最低。

三种搜集数据方法的特点如表 2－3 所示。

表 2－3　搜集数据不同方法的特点

项目	自填式	面访式	电话式
调查时间	慢	中等	快
调查费用	低	高	低
问卷难度	要求容易	可以复杂	要求容易
有形辅助物的使用	中等利用	充分利用	无法利用
调查过程控制	简单	复杂	简单
调查员作用的发挥	无法发挥	充分发挥	一般发挥
回答率	最低	较高	一般

由此可知，没有哪一种方法在所有方面都是最好的，因此，在数据搜集方法的选择中要根据调查所需信息的性质、调查对象的特点、对数据质量和回答率的要求，以及预算费用和时间要求等多方面因素综合而定。也许没有一种方法是完全适用的，这时就要考虑研究人员对数据需求的最主要方面。需要说明的是，各种方法并不是相互排斥的；相反，在许多方面恰恰是相互补充的，因此，在一项调研活动中将各种方法结合起来使用也许是不错的选择。例如，对被选中的调查单位首先采用邮寄问卷方式，让受访者自填，对没有返回问卷的受访者，再进行电话追访或面访。

2.9 什么是实验组和对照组?

- 实验组是指随机抽选的实验对象的子集。在这个子集中，每个单位接受某种特别的处理。
- 对照组中，每个单位不接受实验组成员所接受的某种特别的处理。

2.10 ★★★ 简述抽样误差和非抽样误差

抽样误差是指由抽样的随机性引起的样本结果与总体真值之间的差异。

- 抽样误差只存在于概率抽样中。
- 抽样误差的大小与很多因素有关，
    1. 最主要的是**样本量**大小（样本量越大抽样误差越小）
    2. 总体的**变异性**有关（总体变异性越小抽样误差越小）
    3. **抽样方法**的选择有关（重复抽样与非重复抽样）
    4. **抽样组织方式**不同有关。采取不同的组织方式会有不同的抽样误差，是因为不同的抽样组织所抽中的样本对于总体的代表性是不同的。
- 抽样误差无法避免，但可以计算和控制。

  重复抽样的标准误：

$$\mathrm{SE} = \frac{\sigma}{\sqrt{n}} \tag{1}$$

非重复抽样的标准误：

$$\mathrm{SE} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \tag{2}$$

称$\sqrt{\frac{N-n}{N-1}}$为有限总体矫正（FPC），因此非重复抽样误差小于重复抽样误差。

非抽样误差是相对抽样误差而言的，指除了抽样误差以外由其他原因引起的样本观测结果与总体真值之间的差异。非抽样误差广泛存在于概率抽样与非概率抽样。

①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，我们把这种误差称为抽样框误差。

②回答误差

回答误差是指被调查者在接受调查时给出的回答与真实情况不符。导致回答误差的原因有多种，主要有理解误差、记忆误差和有意识误差。

a.理解误差。不同的被调查者对调查问题的理解不同，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。

b.记忆误差。有时，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，有些事件是按一定周期发生的。

c.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，可能就会提供一个不真实的数字。产生有意识误差的动因大致有两种：一种是调查问题涉及个人隐私，被调查者不愿意告知，所以造假；另一种是受利益驱动，进行数字造假。有意识误差比记忆误差的危害要大。因为记忆误差具有随机性，有些人可能说高了，有些人可能说低了，高低相抵，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。

③无回答误差

无回答误差是指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。无回答也包括那些调查进行时被访者不在家的情况。电话调查中，拨通后没有人接；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，这些都可以视为调查中的无回答误差。

④调查员误差

这是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，调查员的表情变化、语气变化、语速变化都可能对被调查者产生某种影响。

⑤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。

2.11 ★★★ 如何控制误差?

对于抽样误差：抽样误差无法避免，但可以控制和计算。控制抽样误差的主要方式是改变，样本量越大误差越小，利用SE的计算公式，在确定了对误差容忍的限度后可以计算出至少需要多少的样本量。

对于非抽样误差：

2 控制非抽样误差?

01 对于同一个调查问题，可以尝试构造不同的抽样框，对其经过分析比较，挑选出比较好的抽样框，同时广泛搜集信息对抽样框进行改进。例如把抽样框结合起来，以弥补抽样覆盖不全的问题。

02 一份好的调查问卷可以有效的减少调查误差，因此做好问卷设计是减少抽样误差的一个好办法。

03 重要方面是进行调查过程中的质量控制。包括调查员的挑选，培训，督导员的调查专业水平，对调查过程中进行控制的具体措施，对调查结果进行检验评估等。

2.12 ★★★ 影响抽样误差大小的因素有哪些?

抽样误差的大小与很多因素有关，

1. 最主要的是**样本量**大小（样本量越大抽样误差越小）
2. 总体的**变异性**有关（总体变异性越小抽样误差越小）
3. **抽样方法**的选择有关（重复抽样与非重复抽样）
4. **抽样组织方式**不同有关。采取不同的组织方式会有不同的抽样误差，是因为不同的抽样组织所抽中的样本对于总体的代表性不同。

补：抽样调查与典型调查的异同点?

11．抽样调查与典型调查有何异同点？

抽样调查与典型调查的异同点分别为：

（1）相同点

①两种抽样方式都是非全面调查；

②调查单位少，可节省人力、物力、时间；

③灵活性强；

④属于由部分到全面的调查方式

（2）不同点

①定义不同：抽样调查是按照随机原则，从调查总体中抽取部分调查单位进行观察，并根据这一部分调查单位的观察结果，从数量方面推断总体指标的一种非全面调查；典型调查是根据调查目的和要求，在对被研究对象做全面分析的基础上，有意识地从中选择少数具有代表性的典型单位进行深入细致地调查研究，以便认识事物的本质及其规律性的一种非全面调查；

②特点不同：抽样调查的主要特点是按随机原则抽选样本，总体中每一个单位都有一定的概率被抽中，可以用一定的概率来保证将误差控制在规定的范围之内；典型调查的主要特点是调查单位少、机动灵活、典型单位的选择带有一定的主观性、典型单位可以注重于现象数量方面的分析；

③组织形式不同：抽样调查常用的组织方式有简单随机抽样、分层抽样、等距抽样和多阶段抽样等。典型抽样一般有两种方式：一种是一般的典型调查，即对个别典型单位的调查研究；第二种是具有统计特征的划类选点典型调查，即将调查总体划分为若干个类，再从每类中选择若干个典型进行调查，以说明各类的情况。

Ch3 数据的图表展示

3.1 ★★★ 什么是数据的预处理?

数据的预处理是在数据分析之前所做的必要处理，内容包括：

1. 数据的**审核**
2. 数据的**筛选**
3. 数据的**排序**

4. 等等

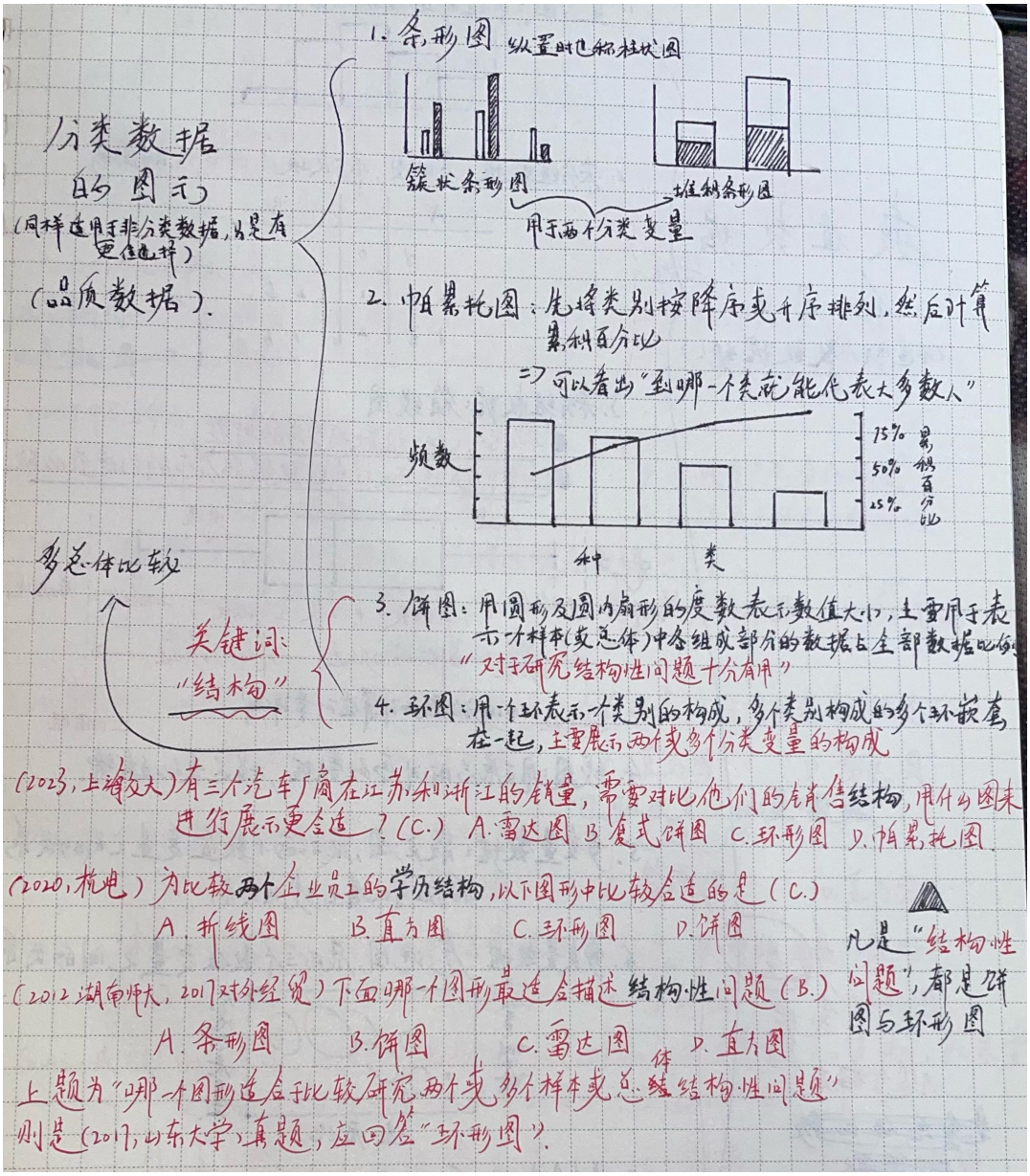
特别地，数据审核中：

- 对原始数据：主要审核完整性与准确性
- 对二手数据：主要检查适用性与时效性

1 数据审核：数据审核检查数据是否存在错误的过程，对于通过调查取得的原始数据，主要从完整性和准确性两个方面去审核。其中完整性主要是检查应调查的单位或者个体是否有遗漏，所有调查项目是否填写齐全等。准确性审核是检查数据是否有错误，是否存在异常值，如果有异常值，正确时则予以保留，若属于记录时的错误，则应分析前予以纠正。

2 数据排序：指的是按一定顺序将数据排列好，以便研究者通过浏览数据发现一些明显的特征或者趋势，找到解决问题的线索，除此以外排序有助于检查纠错，为重新归类或者分组提供了方便。

3.2 简述分类数据可以用哪些图形展示?



3.2 ★★★ 什么是数据分组？以及数据分组步骤

（需要先说明定义，目的以及其两种类型的具体方法）

对分类数据主要是分类整理，对数值数据主要是分组整理。

- 单变量分组：仅针对离散变量且变量值较少时，一个变量即为一组，例如恋爱次数。
- 组距分组：将全部变量依次划分为若干区间，一个区间为一组，适用于连续变量或离散变量且变量值较多时。

整理：数据分组：是指根据研究对象的特征和研究目的，按照一个或几个重要变量，将原始数据划分为性质不同的若干组成部分的一种统计方法。

分组数据：分组后的数据称为分组数据。

数据分组的目的：观察数据分布特征，揭示现象之间存在的差别，保持同步一组内资料的同质性和各组间统计资料的差异性。

分组原则：要正确选择分组标志和正确划分组限，同时要遵循不重不漏的原则，不重是指总体中任一数据都职能出现在其中某一个组，不能出现属于两个或两个以上的组。不漏是指总体中任一单位或数据只能分布在其中某一组，不能出现遗漏。

作用：划分了现象的类型，反映出现象的内部结构，反映了总体各部分之间依存关系等。

具体步骤：

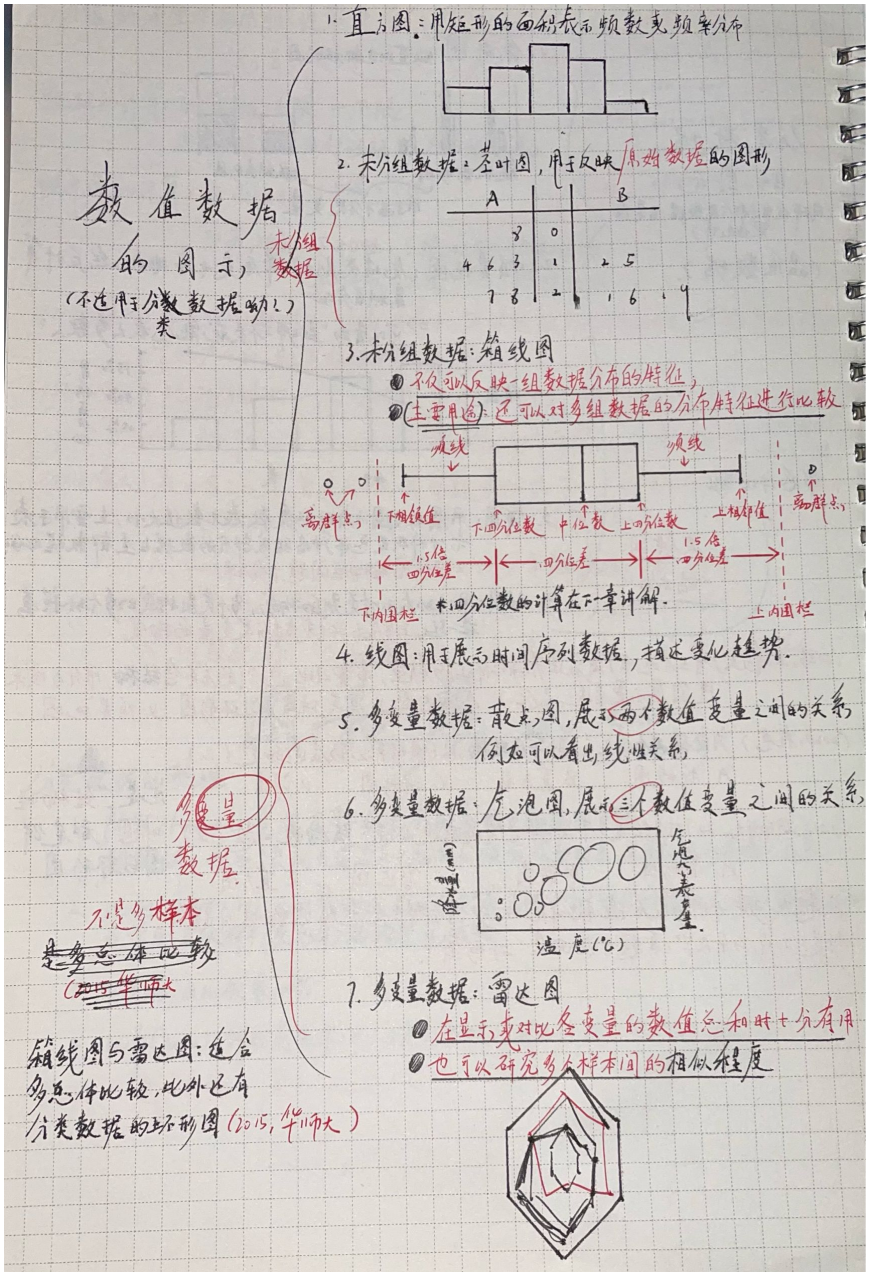
01 确定组数：选择合适的组距可以更好的便于观察分布特征和规律，组数的确定应以能够显示数据的分布特征和规律为目的。一般情况下，K 在 5-15

02 确定组距：组距是一个组上限和下限的差，组距可根据全部数据最大值和最小值来确定，组距等于（最大-最小）/组数，一般取 5 或 10 倍数，且第一组下限略低最小变量值，最后一组上限略高于最大变量值。

03 根据分组整理成频数分布表。

需要注意的是若最大值最小值与其他数据相差悬殊，此时为避免出现空白组或个别极端值被漏掉，第一组和最后一组可以采用 xx 以上，xx 以下开口组。各组组距相等，则称等距分组，各组组距不相等，称为不等距分组。

3.4 简述数值型数据可以用哪些图形展示?



3.5 鉴别图形优劣的准则

一张好图应当：

1. 精心设计，有助于洞察问题的实质
2. 使复杂的观点得到简明、确切、高效的阐述
3. 能在最短的时间内以最少的笔墨提供大量的信息
4. 是多维的
5. 表述数据的真实情况

3.6 统计表的构成有哪些?

表 3.18　1999—2000 年城镇居民家庭抽样调查资料　←表题

项　目		单位	1999 年	2000 年	
行标题	调查户数	户	40044	42220	数字资料
	平均每户家庭人口	人	3.14	3.13	
	平均每户就业人口	人	1.77	1.68	
	平均每户就业面	%	56.43	53.67	
	平均每一就业者负担人数	人	1.77	1.86	
	平均每人全部年收入	元	5888.77	6316.81	
	#可支配收入	元	5854.02	6279.98	
	平均每人消费性支出	元	4615.91	4998.00	

←列标题

注：本表为城镇居民家庭收支抽样调查材料。
资料来源：《中国统计年鉴 2001》，中国统计出版社，2001，第 305 页。　} 附加

3.7 制作统计表时应该注意哪些问题?

首先，要合理安排统计表的结构，例如行标题、列标题、数字资料的位置应安排合理。当然，由于强调的问题不同，行标题和列标题可以互换，但应使统计表的横竖长度比例适当，避免出现过高或过长的表格形式。

其次，表头一般应包括表号、总标题和表中数据的单位等内容。总标题应简明确切地概括出统计表的内容，一般需要表明统计数据的时间(When)、地点(Where)以及何种数据(What)，即标题内容应满足 3W 要求。如果表中的全部数据都是同一计量单位，可放在表的右上角标明，若各指标的计量单位不同，则应放在每个指标后或单列出一列标明。

再次，表中的上下两条横线一般用粗线，中间的其他线要用细线，这样使人看起来清楚、醒目。通常情况下，统计表的左右两边不封口，列标题之间可用竖线分开，而行标题之间通常不必用横线隔开。总之表中尽量少用横竖线。表中的数据一般是右对齐，有小数点时应以小数点对齐，而且小数点的位数应统一。对于没有数字的表格单元，一般用“—”表示，一张填好的统计表不应出现空白单元格。

最后，在使用统计表时，必要时可在表的下方加上注释，特别要注意注明资料来源，以表示对他人劳动成果的尊重，备读者查阅使用。

3.8 ★★★ 简述绘制直方图/茎叶图/箱线图/雷达图/气泡图的步骤

- 直方图：2014年重大432真题参考答案
  1. 通过样本量n的大小**确定组数**的多少，通常取$5 \leqslant k \leqslant 15$
  2. **确定组距**，为方便计算一般取10的倍数
  3. **确定每组的上下限**，做到不重不漏（上组限不在内原则）
  4. **绘制**直角坐标系，用横轴表示分组，纵轴表示频数分布
- 茎叶图（未分组数据）：绘制茎叶图的关键是设计好树茎，通常是以该组数据的高位数值作为树茎，而且树叶上只保留该数值的最后一个数字。树茎一经确定,树叶就自然地长在相应的树茎上了。
- 箱线图（未分组数据）：

1. 首先，找出一组数据的中位数和两个四分位数，并画出箱子
2. 其次，计算出内围栏和相邻值，并画出须线
3. 最后，找出离群点，并在图中单独标出

2019年重大432真题参考答案：

- 先找出一组数据的最大值、最小值、中位数和两个四分位数
- 连接两个四分位数，画出箱子
- 将最大值和最小值与箱子相连，中位数置于箱子的中间

- 雷达图：
  1. 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**
  2. 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p个变量**的坐标轴**
  3. 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个坐标上的点相连线
- 气泡图：
  1. 将**一个变量**放在横轴
  2. 将**另一个变量**放在纵轴
  3. 用气泡的大小表示**第三个变量**的相对数值

补：直方图与条形图的不同

直方图与条形图不同。首先，条形图是用条形的长度（横置时）表示各类别频数的多少，其宽度（表示类别）则是固定的；直方图是用面积表示各组频数的多少，矩形的高度表示每一组的频数或频率，宽度则表示各组的组距，因此其高度与宽度均有意义。其次，由于分组数据具有连续性，直方图的各矩形通常是连续排列，而条形图则是分开排列。最后，条形图主要用于展示分类数据，而直方图则主要用于展示数值型数据。

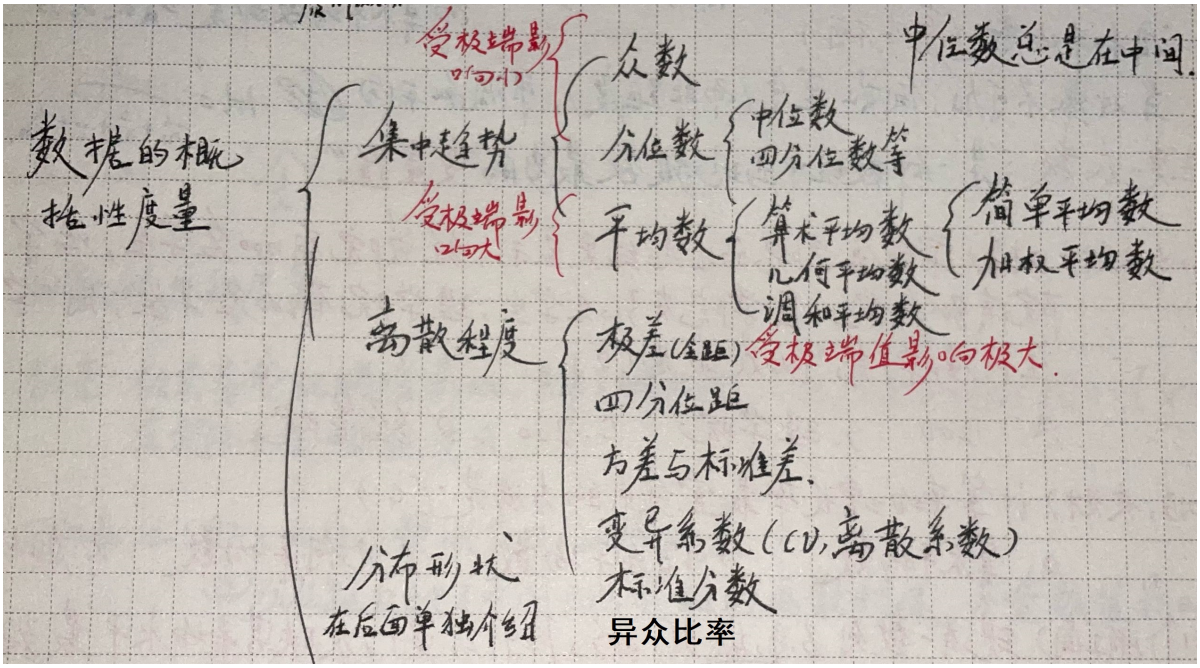
Ch4 数据的概括性度量

4.1 数据的分布特征可以如何测度？（集中，离散，形状）

可以从三个方面进行测度和描述：

1. 分布的集中趋势，反映各数据向其中心值靠拢或据集的程度
2. 分布的离散程度，反映各数据远离其中心值的趋势
3. 分布的形状，反映数据分布的偏态和峰态

4.2 有哪些反映数据集中趋势的度量？（需要熟悉众数，平均数等性质）



4.3 有哪些反映数据离散程度的度量？

见上一问。

4.4 ★★★ 简述众数、平均数、中位数的特点及其应用场合

1 集中趋势：指一组数据向某个中心值靠拢的程度，反映了一组数据中心点的位置所在，它对总体的某一特征具有代表性，表明所研究的理论现象在一定时间，空间条件下的公共性质和一般水平。常用的集中趋势测度值有众数，中位数，分位数和平均数。

2 众数：是一组数据中出现次数最多的变量值，用 Mo 表示。

特点：

01 众数主要用于测度分类数据的集中趋势；

02 数据量较大时，众数才有意义；

03 众数是位置代表值，不受极端值影响；

04 并且一组数据可能有一个也可能有多个众数也可能没有众数。从分布角度来看，众数是一组数据具有明显将集中趋势点的数值。

3 中位数：是一组数据排序以后处于中间位置上的变量值，用 Me 表示。

特点：

01 主要用于测度数值型数据集中趋势，不适用于分类数据；

02 当一组数据偏斜程度比较大时，使用中位数作为其代表值效果比较好。

4 平均数：平均数也称为均值，是一组数据相加以后除以数据的个数得到的结果。

特点：

01 主要适用于数值型数据，不适用于分类数据；

02 当数据呈现对称或接近对称分布时，平均数有较好的代表性。

地位：是集中趋势最广泛的测度值，平均数在统计学中具有重要地位，是进行统计推断和统计分析的基础，平均数是一组数据重心所在，是数据误差相互抵消以后的必然结果，可反映出事物必然性的数量特征。

5 四分位数：也称为四分位点，其特点是不受极端值影响，是一组数据排序以后处于 25%和 75%位置上的值，可通过三个点将数据四等分。

特点：四分位数也是位置代表值，不受极端值的影响。

4.5 ★★★ 离散系数和方差/标准差有何区别?

6 离散程度：反映了各变量值远离其中心值的程度，数据的离散程度越大，集中趋势的测度值对该组数据代表性就越差，离散程度越小，其代表性就越好，描述其所采用的测度值主要

有异众比率，四分位差，方差和标准差，极差，平均差等。

7.异众比率：非众数组的频数占总频数的比例，Vr 表示，公式---。

作用 1：主要用于衡量众数对一组数据的代表程度，异众比率越大，说明非总数组占总频数比例越大，众数代表性越差，异众比率越小，说明非众数组占总频数比重越小，众数代表性越好。

作用 2：主要用于测度分类数据离散程度。

8.四分位差：也称为内距或四分间距，是上四分位数与下四分位数之差，用Q_d表示，反映了中间 50%数据的离散程度，数值越小，说明中间数据越集中，数值越大，说明中间数据越分散，四分位差不受极值影响，主要用于测度顺序数据的离散程度。

9.极差：一组数据最大值与最小值之差，也称为全距，用 R 表示。

特点 1：极差是描述数据离散程度最简单的测度值，计算简单易于理解。

特点 2：但是容易受到极端值影响。

特点 3：同时仅仅利用了一组数据两端信息，不能反映出中间数据的离散状况，因而无法反映数据分散程度。

10.平均差：也称为平均绝对离差 是变量值与其平均数离差绝对值的平均数，用M_d表示，平均差以平均数为中心，反映了每个数据与平均数的平均差异状况，能准确反映一组数据离散状况，平均差越大，说明数据离散程度大，反之说明数据离散程度小，同时计算时带了绝对值，不方便计算，应用少

11.方差和标准差：是各变量值与其平均数离差平方和的平均数，在数学上是通过平方的方法消去离差正负号，然后进行平均，方差的平方根称为标准差，方差可以较好的反映出数据的离散程度，是实际中应用最广泛的离散程度测度值。同时由于方差无量纲，标准差有量纲，因此实际中更常用标准差。

12.标准分数：变量值与其平均数的离差除以标准差以后的值称为标准分数，也称为标准化值或 z 分数，它给出了一组数据中各数值的相对位置，在对多个具有不同量纲的变量进行处理时，通常需标准化处理，标准分数具有平均数为 0，标准差为 1 的特点，实质上是对一组数据进行了线性变换，没有改变一个数据在该组数据中的位置，也没有改变分布的形状，只是将该组数据变为平均数为 0，标准差为 1。

13.离散系数：也称变异系数，是一组数据的标准差与其相应的平均数之比，计算公式--，是测度数据离散程度的统计量，主要用于比较不同样本的离散程度，离散系数越大，说明数据

的离散程度也大，离散系数小，说明数据的离散程度小。

14.异众比率

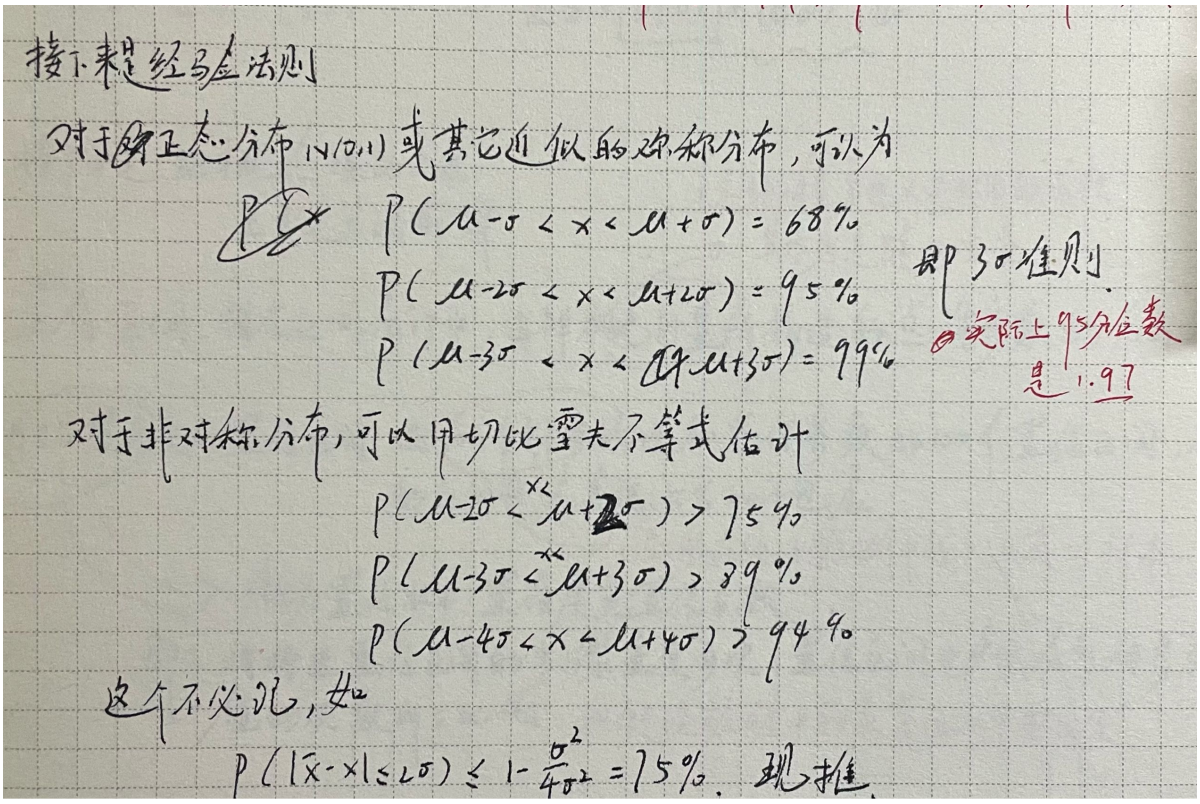
异众比率是指总体中非众数频数与总体全部频数之比，即非众数组的数占总频数的比例，用V_r表示。其计算公式为：

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

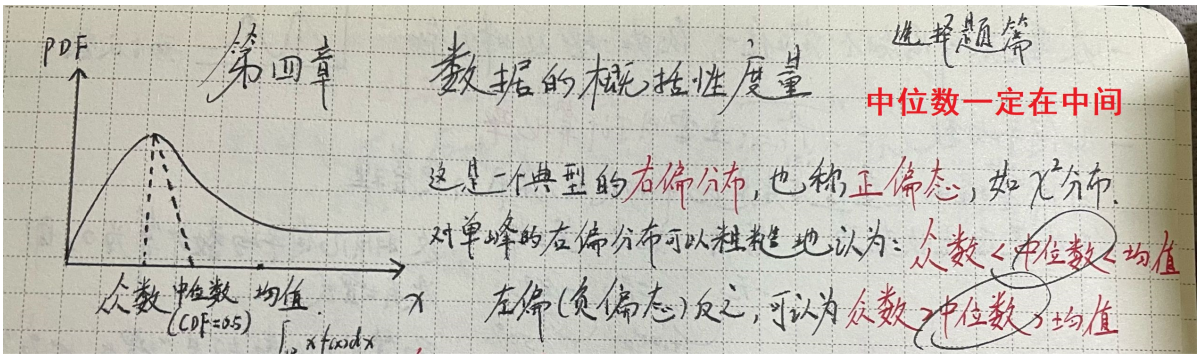
式中，$\sum f_i$为变量值的总频数；f_m为众数组的频数。

异众比率主要用于衡量众数对一组数据的代表程度。异众比率越大，说明非众数组的数占总频数的比重越大，众数的代表性越差；异众比率越小，说明非众数组的频数占总额数的比重越小，众数的代表性越好。异众比率适合测度分类数据的离散程度。

4.6 反映数据相对位置度量的有哪些？（标准分数/经验法则/切比雪夫不等式）



4.7 ★★★ 简述偏度和峰度系数（主要是选择题，但是一定要搞清楚不同数值时是左偏or右偏？尖峰or扁平）



按照贾书的定义，有：

- 记偏度系数为SK，则当$0<|SK|<0.5$称轻微倾斜，$0.5<|SK|<1$称中度倾斜，$|SK|>1$则称严重倾斜
- 标准正态分布的峰度系数为0

Ch6 统计量及其抽样分布

6.1 ★★★ 什么是统计量?

1 统计量：设$X_1,X_2,...,X_n$是从总体抽中取得容量为 n 的样本，如果由此构造一个函数$T(X_1,X_2,...,X_n)$且不依赖于任何未知总体参数，则称其为一个统计量。统计量可以把分散在样本中的信息集中起来，在统计学中，有极其重要地位。针对不同的研究目的，可以构造不同的函数。

6.2 ★★★ 简述中心极限定理

5 中心极限定理：

设从总体均值为 u，方差为一个有限常数的任意总体中，抽取一个样本量为 n 的样本，当 n 充分大时，样本均值 xbar 的抽样分布近似服从均值为 u，方差为 1/n 倍总体方差，通常认为 n 大于等于 30 为大样本，小于 30 为小样本，这是一种经验上的说法。

6.3 ★★★ 简述总体分布，样本分布和抽样分布的含义和特点

8. 总体分布、样本分布和抽样分布是什么

总体分布：总体是我们所关心的若干个元素的集合，总体中每个元素的取值是不同的，这些观察值所形成的分布是总体分布。即总体中各元素的观察值所形成的相对频数分布，称为总体分布。而总体往往无法全部获取，因此总体中相应的参数需要进行估计和推断。

样本分布：从总体中抽取一个容量为n的样本，由这n个观察值所形成的相对频数分布，称为样本分布。当样本容量n逐渐增大时，样本的分布也逐渐接近总体的分布。

抽样分布：样本统计量，如样本均值和样本方差等，均为样本的一个函数。这些样本统计量的概率分布即称为抽样分布，如样本均值和样本方差分布都称为抽样分布。

具体讲：某个统计量的抽样分布，从理论上讲是重复抽取一个容量为 n 的样本时，由该统计量的所有可能取值形成的相对频数分布。

实际上，不可能抽取所有的样本，因此这是一种理论分布。（抽样分布即是用来推断总体中参数的重要科学性保证“工具”）

2015年重大432真题参考答案：举例说明总体分布、样本分布与抽样分布。

- 总体分布：如要调查某校的全部女生的身高，则该校所有女生身高的观测值所形成的相对频数分布即总体分布。
- 样本分布：从该校中抽取一个班的女生，则抽取到的学生身高的观测值所形成的相对频数分布即样本分布。

- 抽样分布：从该校抽取每个班女生的平均身高的所有抽样均值形成的相对频数分布即抽样分布。

Ch7 参数估计

7.1 ★★★ 什么是参数估计?

参数估计是在抽样及抽样分布的基础上，利用样本统计量来估计总体参数，例如用样本均值估计总体均值。

标准答案：参数估计就是用样本统计量去估计总体的参数。

7.2 ★★★ 什么是估计量和估计值

1. 估计量：在参数估计中，用来估计参数的统计量称为估计量，例如样本均值、样本方差等。
2. 估计值：根据一个具体的样本计算出来的估计量的数值。

2021年重大432真题：统计量的含义是（C）

A. 总体参数的名称　　B. 总体参数的具体值

C. 用来估计总体参数的统计量名称　　D. 用来估计总体参数的具体值

7.3 ★★★ 什么是点估计和区间估计?

- 点估计就是用样本估计量$\hat{\theta}$的某个取值直接作为总体参数$\hat{\theta}$的估计值。
- 区间估计是在点估计的基础上给出总体参数的一个区间范围，该区间通常由样本统计量加减估计误差得到。

7.4 ★★★ 如何理解置信区间?

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。

- 置信水平为$1-\alpha$的置信区间，是指**用某种方法构造的所有区间中有$(1-\alpha)\%$的区间包含总体参数的真指，有$\alpha\%$的区间不包含。**
- 置信区间是一个**随机区间。**置信区间会随着样本的不同而不同，因为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，因此抽取不同的样本，就会得到不同的区间。
- 用某样本构造的区间是一个**特定区间**，而不是随机区间。在实际问题中进行估计时通常只抽取一个样本，这时所构造的是与该样本相联系的一定的置信水平下的置信区间，该区间要么包含真值，要么不包含，所以不能说一个特定的区间"有多大的概率包含总体参数"或"真值有多大的概率落在区间内"。

7.5 ★★★ 评价估计量的主要标准有哪些?

最主要有无偏性、有效性、相合性。

如果一个估计量$\hat{\theta}$的数学期望等于被估计的总体参数θ，则称其为无偏估计，即$\mathbb{E}\hat{\theta}=\theta$。

有效性是相对的，是指一个统计量相较于另一统计量有更小的方差。有的总体参数存在一致最小方差无偏估计UMVUE，但如果权衡考虑偏差和方差，则有偏估计中可能有更好的选择。

相合性是对统计量的基本要求，指当样本量$n\to\infty$时，统计量的值应该以某种方式"逼近"被估计的总体参数，换句话说，可以通过大量的重复试验，将误差控制在任意精度内。相合性可以被分为强相合和弱相合等。

7.6 ★★★ 解释独立样本和匹配样本

独立样本是指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。

匹配样本是指一个样本中的数据与另一个样本中的数据**相对应**，例如先指配12个工人用第一种方法组装产品，再让这12个工人用第二种方法组装产品，这样得到的两种方法组装产品的数据就是匹配数据。

Ch8 假设检验

8.1 什么是假设检验?

3 假设检验：假设检验又称为显著性检验，是通过判断样本与样本，样本与总体之间的差异是由于抽样误差引起还是本质差别造成的统计推断方法，首先对总体未知参数或者分布形式提出假设，然后利用样本信息去检验这个假设是否成立的过程。

8.2 ★★★ 参数估计和假设检验有什么异同?

参数估计和假设检验是统计推断的两个组成部分，他们都是利用样本对总体进行的某种推断，但推断的角度不同。

参数估计讨论的是用样本统计量估计总体参数的方法，总体参数θ在估计前是未知的。

而在假设检验中，则是先对θ提出假设，然后利用样本信息去检验这个假设是否成立。

8.3 ★★★ 原假设和备择假设地位有何差异？我们该遵循哪些原则?

2 原假设：原假设又称为零假设，是研究者想搜集数据予以反对的假设。通常是检验中把原有的，传统的一些观点或结论在一次试验中，原假设是受到保护的的。用 H0 表示。

备择假设：备择假设是在一次试验中不容易发生的事件，与原假设互斥，进行检验时，一般是把希望证明的命题放在备择假设上。用 H1 表示。

2014年重大432真题参考答案：

1. 一般将研究者想予以证明的假设作为H_1，将想予以驳斥的假设作为H_0
2. 为方便检验，一般将清晰的、即把 “=” 关系置于H_0
3. 先确定H_1，再确定H_0
4. 假设检验的主要目的是搜集证据拒绝H_0

> 假设检验是在以H_0为真的前提下进行的，如果能够拒绝H_0，则接受H_1
>
> - 最好把常识和过去的真理作为H_0，把希望推翻的结论作为H_0
> - 把想要证明的结果置于H_1

8.4 ★★★ 简述假设检验的两类错误，并说明为什么要控制 I 类错误的概率

7 两类错误及两类错误的关系？

关系：对于一定的样本量 n，不能同时使得犯两类错误的概率都很小，如果减小 α 错误，会增大犯 β 错误的概率，如果减小 β 错误就会增大犯 α 错误的机会。通过增大样本量可以使得犯两类错误的概率减小，但是样本量必须是有限的，否则也就失去了调查意义。

一般来说，哪一类错误带来的后果越严重，在假设检验中优先作为控制目标。通常是控制 α 错误为原则 理由如下：

01 从实用的观点来看，原假设常是原有的观点或结论，常常是明确的，备择假设是什么常常是模糊的。对于一个含义清楚和一个含义模糊的假设，我们更愿意接受前者，即更关心 H0 为真的，我们却拒绝了，犯这种错误的可能性有多大。

02 遵循统一的原则，讨论问题比较方便。

8.5 ★★★ 什么是显著性水平与统计显著?

3 如何理解显著性水平的含义

定义：通常把 a 称为显著性水平，含义是当原假设正确时，却被拒绝的概率或风险。是在检验之前由人们根据检验的要求确定好的，常取 α 等于 0.01，0.05，这表示当接受原假设时，其正确的概率为 1-α。

显著性水平取α，意味着在原假设成立时，如果事件的发生概率小于α则认为原假设不成立。换言之，我们有$1-\alpha$的把握拒绝原假设。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。

（1）在假设检验中，拒绝原假设称样本结果在“统计上是显著的”；不拒绝原假设则称结果是“统计上不显著的”。 也即认为原假设与样本观测结果的差异显著，这个差异大到了认为是存在实质性或系统性因素造成，而不仅仅是抽样误差。

（2）“显著的”在这里是指“非偶然的”，它表示这样的样本结果不是偶然得到的。

（3）在显著和不显著之间没有清楚的界限，只是P值越来越小时，我们有越来越强的证据而已。

8.6 ★★★ 假设检验的基本思路和步骤

思路：利用小概率原理与反证法，其中小概率原理是指发生概率很小的随机事件在一次实验中几乎不可能发生。

步骤：

1. 提出**原假设**H_0与**备择假设**H_1
2. 从研究总体中**抽取一个随机样本**
3. 构造**检验统计量**，并根据样本计算出统计量的数值
4. 确定**显著性水平**，计算拒绝域（计算临界值）

5. **统计决策**，如果落入拒绝域则拒绝原假设，反之不能拒绝原假设

8.7 ★★★ P值的含义、优点及影响因素

(不要与α、β弄混淆了)

2 假设检验中 P 值含义以及检验步骤，优点

定义：P 值就是当原假设正确时，出现的样本观测结果或者更极端结果出现的概率，是计算出的一个反映观察到样本数据与实际原假设之间不一致的概率值。若 P 值很小，说明这种情况发生的概率很小，如果发生了，根据小概率原理，有充足的理由拒绝原假设，P 值越小，拒绝原假设的理由越充分。

影响 P 值的因素：一是样本数据与原假设之间的差异，二是样本量，三是被假设总体参数分布。

优点：P 值是一个计算出的，反映样本观测数据与原假设之间不一致的概率值。提供了更多的信息，同时，由于用传统的拒绝域来进行决策时，对于不同样本，所面临的风险度都是 α，实际中，不同样本面临的风险度是不一样的，P 值可以精确的反映出不同样本面临的风险度。进行决策时，将 P 值与给定的显著性水平进行对比，若 P 小于 α，则拒绝原假设；P 大于 α，则没有充足的理由拒绝原假设。

8.8 ★★★ 什么是统计显著?

(总结："统计显著" 就是 "结果不偶然")

3 如何理解显著性水平的含义

定义：通常把 a 称为显著性水平，含义是当原假设正确时，却被拒绝的概率或风险。是在检验之前由人们根据检验的要求确定好的，常取 α 等于 0.01，0.05，这表示当接受原假设时，其正确的概率为 1-α。

显著性水平取α，意味着在原假设成立时，如果事件的发生概率小于α则认为原假设不成立。换言之，我们有$1-\alpha$的把握拒绝原假设。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。

（1）在假设检验中，拒绝原假设称样本结果在"统计上是显著的"；不拒绝原假设则称结果是"统计上不显著的"。 也即认为原假设与样本观测结果的差异显著，这个差异大到了认为是存在实质性或系统性因素造成，而不仅仅是抽样误差。

（2）"显著的"在这里是指"非偶然的"，它表示这样的样本结果不是偶然得到的。

（3）在显著和不显著之间没有清楚的界限，只是P值越来越小时，我们有越来越强的证据而已。

8.9 ★★★ 显著性水平和P值有何区别?

通常把 a 称为显著性水平，含义是当原假设正确时，却被拒绝的概率或风险。

是在检验之前由人们根据检验的要求确定好的，常取 α 等于 0.01，0.05，这表示当拒绝原假设时，其错误的概率不超过 α。

P 值就是当原假设正确时，出现的样本观测结果或者更极端结果出现的概率，是计算出的一个反映观察到样本数据与实际原假设之间不一致的概率值。若 P 值很小，说明这种情况发生的概率很小，如果发生了，根据小概率原理，有充足的理由拒绝原假设，P 值越小，拒绝原假设的理由越充分。

联系：

在假设检验中，做出统计决策时，一般将给定的显著性水平 α 与 P 值比较，p 小于 α，应该拒绝原假设，P 大于 α 没有足够的理由说明原假设错误，故不能拒绝。

区别：

01 显著性水平是人们在检验之前根据检验要求确定的值，一般取 0.05 或者 0.01；P 值是通过计算得到的，反映了观察到的数据与原假设之间不一致的概率值。

02 在统计决策时，在给定的显著性水平下，根据抽样分布的出原假设成立时的临界值，由此构成的拒绝域在统计决策时，对于不同的样本，其面临拒绝原假设的风险是一致的，而实际中，不同样本决策时面临犯错误风险往往不一致，P 值精确的反映决策的风险度。

补：假设检验的原理

5 假设检验的原理

定义：假设检验又称为显著性检验，是通过判断样本与样本，样本与总体之间的差异是由于抽样误差引起还是本质差别造成的统计推断方法，首先对总体未知参数或者分布形式提出假设，然后利用样本信息去检验这个假设是否成立的过程。

假设检验的基本原理：概率性质反证法，即推断依据是小概率原理，所谓小概率原理是指一次随机试验中，小概率事件是几乎不可能发生的。

在假设检验中，根据样本资料计算出的统计量值落入拒绝域的概率很小为 0.05 或者 0.01，若在一次试验中，统计量落入了拒绝域，此时有充足理由怀疑原假设和样本资料之间的差异不仅仅由于抽样误差引起，而是存在本质差异，也即拒绝原假设。

例如，10000 件产品中，仅有一件次品，在一次随机试验中，抽到次品概率很小，若真的抽到了，就可以断定此次品数应该不少，否则就不会轻易抽到，也即有充足理由怀疑原假设正确性。

Ch9 分类数据分析

9.1 ★★★ 什么是拟合优度检验，简述其步骤

拟合优度检验是利用χ^2统计量进行统计量显著性检验的重要内容之一。

根据总体分布状况，计算出分类变量中各类别的期望频数，后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性差异，从而达到对分类变量进行分析的目的。

> $Pearson\ \chi^2$检验：记一共有j个类，在原假设H_0：“类A_i占比p_i成立”的条件下，“$\sum$差的平方的加权 $=\sum$差的平方 $\times \frac{1}{\text{期望频数}}$”的渐进分布为自由度为$j-1$的$\chi^2$分布，即：
>
> $$\chi^2=\sum_{i=1}^{j}\frac{(n_i-np_i)}{np_i}\overset{L}{\sim}\chi^2(j-1) \tag{3}$$
>
> 若还含k个未知参数，则用极大似然估计替代未知参数，χ^2分布的自由度变为$j-k-1$。

$$\chi^2=\sum\frac{(f_o-f_e)^2}{f_e} \tag{4}$$

其中f_o是类别观察到的频数，f_e是类别的期望频数。

步骤：

1. 提出原假设与备择假设
2. 根据样本计算χ^2检验统计量的值$\chi^2=\sum\frac{(f_o-f_e)^2}{f_e}$
3. 计算拒绝域，在H_0成立的条件下$\chi^2\sim\chi^2(R-1)$，其中R是分类数
4. 统计决策

9.2 ★★★ 什么是列联表的独立性检验?

独立性检验就是分析列联表中的行变量和列变量是否相互独立。

步骤和拟合优度检验相同，不过令f_e指定为$p_{i+}p_{+j}$。

9.3 ★★★ 简述φ系数，C系数以及V系数的定义及其特点

- φ系数

  $$\varphi=\sqrt{\frac{\chi^2}{n}} \tag{5}$$

  φ系数是描述2×2列联表数据相关程度最常用的一种相关系数，取值范围为$0\sim1$之间，φ的值越大，说明两个变量间的相关程度越高，当$\varphi=1$则说明变量间完全相关。

  但是φ系数不适合行数或列数大于2的列联表，此时φ值没有上限，用φ测量相关程度就不够清晰。

- C系数

  $$C=\sqrt{\frac{\chi^2}{\chi^2+n}} \tag{6}$$

  C系数主要用于行数或列数大于2的列联表，当两变量独立时$C=0$。C系数的特点是不可能大于1，且其可能的最大值依赖于列联表的行数与列数，并且会随着行数或列数的增大而增大，所以对不同行列计算出的C系数不方便直接比较，这是其局限。

- V系数

$$V=\sqrt{\frac{\chi^2}{n\times\min\{R-1,C-1\}}} \tag{7}$$

当两变量完全独立时，$V=0$；当两变量完全相关时，$V=1$；当列联表中有一维为2时，$V=\varphi$。

Ch10 方差分析

10.1 什么是方差分析?

方差分析（ANOVA）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。

1 方差分析/单因素方差分析

用于检验两个即两个以上的样本均值差别的显著性检验，通过检验各总体均值是否相等考判断分类型自变量对数值型因变量是否有显著影响。根据所分析的分类型自变量多少，可分为单因素方差分析和双因素方差分析，当仅涉及一个分类型自变量时称为单因素方差分析。

10.2 ★★★ 简述方差分析的结构

10.2.1 数据结构

进行单因素方差分析时，需要得到下面的数据结构，如表 10－2 所示。

表 10－2 单因素方差分析的数据结构

观测值（j）	因素（i）			
	A_1	A_2	…	A_k
1	x_{11}	x_{21}	…	x_{k1}
2	x_{12}	x_{22}	…	x_{k2}
⋮	⋮	⋮	⋮	⋮
n	x_{1n}	x_{2n}	…	x_{kn}

水平

观测值

为叙述方便，在单因素方差分析中，用 A 表示因素，因素的 k 个水平（总体）分别用 A_1，A_2，…，A_k 表示，每个观测值用 x_{ij}（$i=1$，2，…，k；$j=1$，2，…，n）表示，即 x_{ij} 表示第 i 个水平（总体）的第 j 个观测值。例如，x_{21} 表示第二个水平的第一个观测值。其中，从不同水平中所抽取的样本量可以相等，也可以不相等。

10.3 ★★★ 方差分析的基本思想是什么?

3 方差分析的基本原理?

定义：用于检验两个即两个以上的样本均值差别的显著性检验，通过检验各总体均值是否相等考判断分类型自变量对数值型因变量是否有显著影响，根据所分析的分类型自变量多少，可分为单因素方差分析和双因素方差分析，原理基本如下：

01 误差分解：组内误差是来自水平内部的误差，仅包含随机性导致的误差，反映其大小的平方和称为组内平方和，记为 SSE。组间平方和来自不同水平之间的数据误差，包含系统误差和随机误差。反映其大小的平方和记为 SSA，总平方和是反映全部数据误差大小的平方和，记为 SST。

02 原理：如果组间误差只包含随机误差，没有系统误差，那么组间误差和组内误差经过平均以后的数值均方误差就会很接近，他们的比值就会很接近 1；反之如果组间误差包含了系统误差，那么组间误差经过平均以后的数值就会大于 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显著影响。

10.4 ★★★ 方差分析的基本假定

1. 每个总体都应服从**正态分布**。也就是说，对于因素的每一个水平，观测值都应该是来自正态分布的简单随机样本。
2. 各个总体的**方差σ^2必须相同**。也就是说，各组实验数据是从具有相同方差的正态分布中抽取的。
3. 观测值是**独立**的。

10.5 ★★★ 简述方差分析的基本过程（以单因素方差分析为例，双因素要了解即可）

1. 提出假设

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

   H_1：$\mu_i\ (i = 1, 2, \cdots, k)$不全相等，即自变量对因变量有显著影响

2. 构造检验统计量

   在H_0成立的条件下，有$F = \dfrac{MSA}{MSE} = \dfrac{\frac{SSA}{k-1}}{\frac{SSE}{n-k}} \sim F(k-1, n-k)$

单因子方差分析表

来源	平方和	自由度	均方	F比	p值
因子	S_A	$f_A = r - 1$	$MS_A = \frac{S_A}{f_A}$	$F = \frac{MS_A}{MS_e}$	p-value
误差	S_e	$f_e = n - r$	$MS_e = \frac{S_e}{f_e}$		
总和	S_T	$f_T = n - 1$			

3. 作出统计决策，若$F > F_\alpha$则拒绝原假设，反之则不能拒绝原假设

10.6 ★★★ 方差分析中多重比较有何作用?

通过应用最小显著差异法，可判断研究的几个总体中，究竟是哪两个均值不同，即通过均值之间的配对比较来进一步检验到底哪些均值间存在差异。

10.7 ★★★ 为什么检验多个总体均值是否相等不采用两两比较的t检验而采用方差分析的方法?

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？

（1）假设要检验n（$n>2$）个总体的均值是否相等，若选择作两两比较，则需要检验C_n^2次，而若采用方差分析，一次就可以完成，省去了繁琐的两两比较过程。

（2）进行两两比较要做C_n^2次检验，若每次检验的犯错的概率为ε，则多次检验累积起来的犯错概率将达到$1-(1-\varepsilon)^{C_n^2}$，这样总的犯错误的概率就会变大。

基于以上两个原因，在检验多个总体均值是否相等时，使用方差分析的方法相对较好。

总结：因为 ① **效率更高**，② **犯错的概率会增大。**

10.8 方差分析中R^2的含义和作用（单因素和双因素）

单因素方差分析：

8.单因素方差分析中的 R^2 有什么含义？给出它发生作用的基本原理。

$$R^2=\frac{SSA}{SST}$$

它表示自变量对因变量的影响效应占总效应的比例，其平方根可以用来测量两个变量之间的关系强度。发生作用的基本原理：

（1）SST（总平方和）：对全部数据差程度的度量，反映了自变量和残差变量的共同影响，等于自变量效应加残差效应。

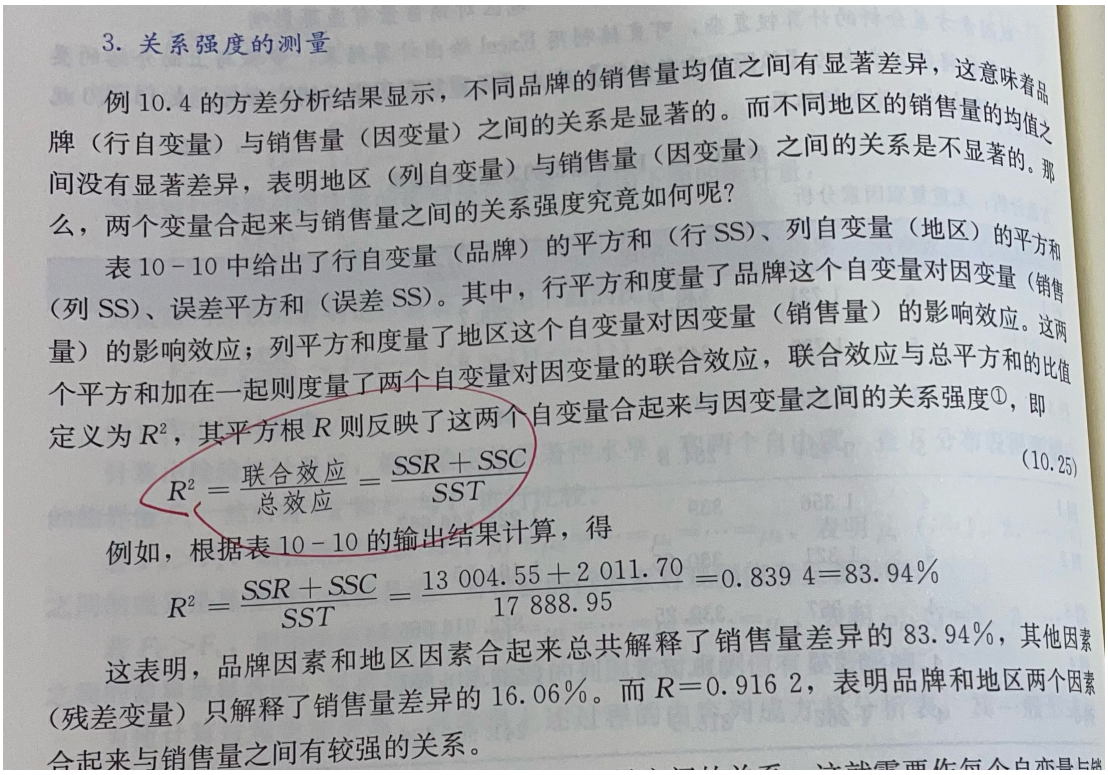
（2）SSA（组间平方和）：是各组均值与总均值的误差平方和，是对随机误差和系统误差大小的度量，反映了自变量对因变量的影响，也称为自变量效应或因子效应。

（3）SSE（组内平方和）：是每个水平或组的各样本数据与其组均值的误差平方和，是对随机误差的度量，反映了每个样本观测值的离散状况，是除自变量对因变量的影响之外，其他因素对因变量的影响，也称为残差变量，引起残差效应。

（4）SST（总平方和）$=SSA$（组间平方和）$+SSE$（组内平方和）

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，大的越多，表明他们之间的关系就越强：反之，当组间平方和比组内平方和小时，意味着两个变，量之间的关系不显著，小的越多，表明他们之间的关系就越弱。所以可以用组间平方和（SSA）占总平方和（SST）的比例大小来测量两个变量之间的关系强度。

多因素方差分析：



3. 关系强度的测量

例 10.4 的方差分析结果显示，不同品牌的销售量均值之间有显著差异，这意味着品牌（行自变量）与销售量（因变量）之间的关系是显著的。而不同地区的销售量的均值之间没有显著差异，表明地区（列自变量）与销售量（因变量）之间的关系是不显著的。那么，两个变量合起来与销售量之间的关系强度究竟如何呢？

表 10－10 中给出了行自变量（品牌）的平方和（行 SS）、列自变量（地区）的平方和（列 SS）、误差平方和（误差 SS）。其中，行平方和度量了品牌这个自变量对因变量（销售量）的影响效应；列平方和度量了地区这个自变量对因变量（销售量）的影响效应。这两个平方和加在一起则度量了两个自变量对因变量的联合效应，联合效应与总平方和的比值定义为 R^2，其平方根 R 则反映了这两个自变量合起来与因变量之间的关系强度①，即

$$R^2 = \frac{\text{联合效应}}{\text{总效应}} = \frac{SSR + SSC}{SST} \tag{10.25}$$

例如，根据表 10－10 的输出结果计算，得

$$R^2 = \frac{SSR + SSC}{SST} = \frac{13\,004.55 + 2\,011.70}{17\,888.95} = 0.839\,4 = 83.94\%$$

这表明，品牌因素和地区因素合起来总共解释了销售量差异的 83.94%，其他因素（残差变量）只解释了销售量差异的 16.06%。而 $R = 0.916\,2$，表明品牌和地区两个因素合起来与销售量之间有较强的关系。

10.9 为什么双因素方差分析优于分别做单因素方差分析?

单因素方差分析只是考虑一个分类型变量对数值型因变量的影响，而在实际问题的研究中，有时需要考虑几个因素对试验结果的影响。

1. 单因素方差分析主要是为了实现三个或更多的平均值之间的平等检验。双因素方差分析是为了评估两个自变量与因变量的相互关系。
2. 单因素方差分析只涉及一个因素或自变量，而双因素方差分析则有两个自变量。
3. 在单因素方差分析中，所分析的一个因素或自变量有三个或多个分类组。双因素方差分析则是对两个因素的多个组进行比较。
4. 单因素方差分析只需要满足实验设计的两个原则，即重复和随机。而双因素方差分析则满足实验设计的所有三个原则，即重复、随机和局部控制。

Ch11 一元线性回归

11.1 ★★★ 什么是相关分析?

相关关系：变量之间存在的不确定的数量关系。相关关系的特点是一个变量的取值并不是由另一个变量唯一确定，即两个变量间的取值不是一一对应的。

相关分析就是对两个变量之间线性关系的描述与度量，他要解决的问题包括：

- 变量之间是否存在关系?
- 如果存在关系，他们之间是什么样的关系?
- 变量之间的关系强度如何?
- 样本所反映的变量之间的关系是否能代表总体变量之间的关系?

11.2 相关分析中有哪些基本假定?

为了解决上述问题，在进行相关分析时，对总体主要有以下两个假定：

- 第一，两个变量之间是线性关系
- 第二，两个变量都是随机变量

11.3 ★★★ 简述相关系数的性质

相关系数是根据样本数据计算的度量两个变量之间线性关系强度的统计量。

$$\rho=\frac{\operatorname{Cov}(X, Y)}{\sqrt{D_X D_Y}},\ r=\hat{\rho}=\frac{l_{xy}}{\sqrt{l_{xx} l_{yy}}}\ \ (\text{一元}) \tag{8}$$

相关系数r的性质：

1. r的**取值范围**是$[-1,1]$，若$0<r \leqslant 1$则表明x与y之间存在正线性相关关系，若$-1 \leqslant r<0$则表明x与y之间存在负线性相关关系。若$r=1$则表明x与y之间为完全正线性相关关系，若$r=-1$则表明x与y之间为完全负线性相关关系，即$|r|=1$时，y的值完全依赖于x，二者之间为函数关系。当$r=0$，y的取值与x无关，二者之间不存在线性相关关系。
2. r具有**对称性**。有$r_{x,y}=r_{y,x}$
3. r的**数值大小**与x和y的**原点及尺度无关**。改变x和y的数据原点及计量尺度，并不改变r的数值大小。
4. r仅仅是x与y之间线性关系的一个度量，**不能用于描述非线性关系**。即$r=0$只表示两个变量之间不存在线性关系，并不说明变量之间没有任何关系，他们之间可能存在非线性关系。变量之间非线性相关程度相当大时，可能会导致$r=0$，因此$r=0$或r很小的时候，不能轻易得出两个变量之间不存在相关关系的结论。
5. r只是两个变量之间线性关系的一个度量，**并不意味**着x与y一定有**因果关系**。

当$|r| \geqslant 0.8$时，可视为高度相关；当$0.5 \leqslant|r|<0.8$时，可视为中度相关；当$0.3 \leqslant|r|<0.5$时，可视为帝都相关；当$|r|<0.3$时，可视为不相关。

11.4 ★★★ 为什么要对相关系数进行显著性检验？并简述其过程

由于样本是随机抽取的，不同的样本计算出的相关系数是不一样的，因此给定一个样本，就有一个相关系数与其对应，即r是一个随机变量。同时，由于抽样波动的影响，此时必须要对样本系数可靠性进行显著性检验，以判断样本所反映的信息能否代表总体变量之间的关系。

相关系数的显著性检验步骤：

1. 选取统计量：通常选取t检验，因为r的正态分布假设风险较大
2. 提出假设：$H_0:\rho=0\ \ vs\ \ H_1=\rho \neq 0$
3. 计算统计量：在H_0成立的条件下，$t=|r| \sqrt{\frac{n-2}{1-r^2}} \sim t(n-2)$，根据样本统计量的抽样分布计算出拒绝域
4. 统计决策

11.5 简述相关系数r的抽样分布

5.样本相关系数 r 的抽样分布

r 的抽样分布随着总体的相关系数 ρ 和样本量 n 的大小而变化，当样本数据来自正态总体时，随着 n 的增大：r 的抽样分布趋于正态分布，尤其是当总体相关系数很小或者接近于 0 时。当 ρ 为较大的正值时，r 呈现左偏分布；当 ρ 为较大的负值时，r 呈现右偏分布。只有当 ρ 接近于 0，而样本量 n 很大时，才能认为 r 是接近正态分布的随机变量。

11.6 ★★★ 一元线性回归的基本假设

首先是对模型和抽样的假定：

1. **线性关系**：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

接下来是对误差的假定：

3. **零均值**：误差项ε期望为0（弱外生性）
4. **同方差**：$\forall x$，$\mathrm{Var}(\varepsilon)=\sigma^2$均相同（球形扰动项）
5. **正态**且**不相关**：ε是一个独立服从正态分布的随机变量，且相互独立（球形扰动项与正态假设）
6. 此外，还有求无多重共线性等

11.7 ★★★ 参数最小二乘估计的基本原理

10.2.2 参数的最小二乘估计

对于第 i 个 x 值，估计的回归方程可表示为

$$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 x_i。 \tag{10.7}$$

对于 x 和 y 的 n 对观察值，用于描述其关系的直线有多条，究竟用哪条直线来代表两个变量之间的关系，需要有一个明确的原则。我们自然会想到距离各观测点最近的一条直线，用它来代表 x 与 y 之间的关系与实际数据的误差比其他任何直线都小。根据这一思想确定直线中未知常数 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的方法称为最小二乘法(method of least squares)。

▶定义 10.9 使因变量的观察值 y_i 与估计值 $\hat{y}_i$ 之间的离差平均和达到最小来求得 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的方法，称为最小二乘法(method of least squares)。

11.8 解释总平方和、回归平方和、残差平方和

从图 10.7 可以看出，每个观测点的离差都可以分解为

$$y-\bar{y}=(y-\hat{y})+(\hat{y}-\bar{y})。 \tag{10.13}$$

将式(10.13)两边平方，并对所有 n 个点求和，有

$$\sum(y_i-\bar{y})^2=\sum(y_i-\hat{y}_i)^2+\sum(\hat{y}_i-\bar{y})^2+2\sum(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})。 \tag{10.14}$$

可以证明，$\sum(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})=0$，因此有

$$\sum(y_i-\bar{y})^2=\sum(y_i-\hat{y}_i)^2+\sum(\hat{y}_i-\bar{y})^2。 \tag{10.15}$$

即总平方和 SST 可分解为两部分，其中 $\sum(\hat{y}_i-\bar{y})^2$ 是回归值 $\hat{y}_i$ 与均值 $\bar{y}$ 的离差平方和，根据估计的回归方程，估计值 $\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_i$，因此可以把 $\hat{y}_i-\bar{y}$ 看作是由于自变量 x 的变化引起的 y 的变化，而其平方和 $\sum(\hat{y}-\bar{y})^2$ 则反映了 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，它是可以由回归直线来解释的 y_i 变差部分，称为回归平方和，记为 SSR。另一部分 $\sum(y_i-\hat{y}_i)^2$ 是各实际观测点与回归值的残差 $y_i-\hat{y}_i$ 的平方和，它是除了 x 对 y 的线性影响之外的其他因素对 y 变差的作用，是不能由回归直线来解释的 y_i 变差部分，称为残差平方和，记为 SSE。三个平方和的关系为

总平方和＝回归平方和＋残差平方和，

即

$$SST=SSR+SSE。 \tag{10.16}$$

从图 10.7 可以直观地看出，回归直线拟合的好坏取决于 SSR 及 SSE 的大小，或者说取决于回归平方和 SSR 占总平方和 SST 比例 SSR/SST 的大小。各观测点越是靠近直线，SSR/SST 则越大，直线拟合得越好。

11.9 ★★★ 什么是判定系数?

▶定义 10.10　回归平方和占总平方和的比例，称为判定系数(coefficient of determination)，记为 R^2。

R^2 的计算公式为

$$R^2=\frac{SSR}{SST}=\frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\bar{y})^2}=1-\frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2}。 \tag{10.17}$$

判定系数 R^2 测度了回归直线对观测数据的拟合程度。如果所有观测点都落在直线上，残差平方和 $SSE=0$，$R^2=1$，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，此时 $\hat{y}=\bar{y}$，则 $R^2=0$。可见 R^2 的取值范围是[0,1]。R^2 越接近于 1，表明回归平方和占总平方和的比例越大，回归直线与各观测点越接近，用 x 的变化来解释 y 值变差的部分就越多，回归直线的拟合程度就越好；反之，R^2 越接近于 0，回归直线的拟合程度就越差。

11.10 ★★★ 相关系数和判定系数的关系

可见在一元线性回归中，相关系数 r 实际上是判定系数的平方根。这一结论不仅可以使我们能由相关系数直接计算判定系数 R^2，也可以使我们进一步理解相关系数的意义。相关系数 r 与回归系数 $\hat{\beta}_1$ 的正负号是相同的，实际上，相关系数 r 也从另一个角度说明了回归直线的拟合优度。$|r|$ 越接近 1，表明回归直线对观测数据的拟合程度就越高。但用 r 说明回归直线的拟合优度需要慎重，因为 r 的值总是大于 R^2 的值（除非 $r=0$ 或 $|r|=1$）。比如，当 $r=0.5$ 时，表面上看似乎有一半的相关了，但 $R^2=0.25$，实际上我们只能解释总变差的 25%。$r=0.7$ 才能解释近一半的变差，$r<0.3$ 意味着只有很少一部分变差可由回归直线来解释。

11.11 ★★★ 简述回归分析中回归估计标准误差的计算及含义

▶定义 10.11　均方残差（MSE）的平方根，称为估计量的标准差或标准误差（standard error of estimate），用 s_y 来表示。

估计标准误差是对各观察点在直线周围分散程度的一个度量值，它是对误差项 ε 的标准差 σ 的估计。其计算公式为

$$s_y=\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n-2}}=\sqrt{\frac{SSE}{n-2}}=\sqrt{MSE}。\tag{10.20}$$

估计标准误差 s_y 可以看作是在排除了 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看，它反映了用估计的回归方程预测因变量 y 时预测误差的大小。若各观测点越靠近直线，s_y 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；若各观测点全部落在直线上，则 $s_y=0$。此时用自变量来预测因变量时是没有误差的。可见 s_y 也从另一个角度说明了回归直线的拟合优度。

从式(10.20)容易看出，回归直线是对 n 个观测点拟合的所有直线中，估计标准误差最小的一条直线，因为回归直线使 $\sum(y_i-\hat{y}_i)^2$ 为最小确定的。

11.12 ★★★ 一元线性回归方程中线性关系的检验及其步骤

1　线性关系的检验

线性关系的检验是检验自变量 x 和因变量 y 之间的线性关系是否显著，或者说，它们之间能否用一个线性模型 $y=\beta_0+\beta_1 x+\varepsilon$ 来表示。为检验两个变量之间的线性关系是否显著，我们需要构造用于检验的一个统计量。该统计量的构造是以回归平方和(SSR)以及残差平方和(SSE)为基础的。将 SSR 除以其相应的自由度(自变量的个数 p，一元线性回归中自由度为 1)后的结果称为均方回归，记为 MSR；将 SSE 除以其相应的自由度($n-p-1$，一元线性回归中自由度为 $n-2$)后的结果称为均方残差，记为 MSE。如果原假设成立($H_0:\beta_1=0$，两个变量之间的线性关系不显著)，则比值 MSR/MSE 的抽样分布服从分子自由度为 1、分母自由度为 $n-2$ 的 F 分布，即

$$F=\frac{SSR/1}{SSE/(n-2)}=\frac{MSR}{MSE}\sim F(1,n-2)。\qquad(10.21)$$

所以当原假设 $H_0:\beta_1=0$ 成立时，MSR/MSE 的值应接近 1，但如果原假设 $H_0:\beta_1=0$ 不成立，MSR/MSE 的值将变得无穷大。因此，较大的 MSR/MSE 值将导致拒绝原假设 H_0，我们就可以断定变量 x 与 y 之间存在着显著的线性关系。线性关系检验的具体步骤如下：

第 1 步：提出假设

$H_0:\beta_1=0$　两个变量之间的线性关系不显著；

第 2 步：计算检验统计量 F

$$F=\frac{SSR/1}{SSE/(n-2)}=\frac{MSR}{MSE};$$

第 3 步：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 F_α。若 $F>F_\alpha$，拒绝 H_0，表明两个变量之间的线性关系是显著的；若 $F<F_\alpha$，不拒绝 H_0，表明两个变量之间的线性关系不显著。

11.13 ★★★ 如何评价回归分析结果?

（1）所估计的回归系数 $\hat{\beta}_1$ 的符号是否与理论或事先预期的相一致。例如，在不良贷款与贷款余额的回归中，贷款余额越多，不良贷款也可能越多；也就是说，回归系数 $\hat{\beta}_1$ 的值应该是正的，在上面建立的回归方程中，得到的回归系数 $\hat{\beta}_1=0.037\ 895$，为正值。

（2）如果理论上认为 y 与 x 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。例如，在不良贷款与贷款余额的回归中，二者之间为正的线性关系，而且对回归系数 $\hat{\beta}_1$ 的 t 检验结果表明，二者之间的线性关系是统计上显著的。

（3）回归模型在多大程度上解释了因变量 y 取值的差异？可以用判定系数 R^2 来回答这一问题。例如，在不良贷款与贷款余额的回归中，得到的 $R^2=71.16\%$，解释了不良贷款变差的 2/3 以上，说明拟合的效果还算不错。

（4）考察关于误差项 ε 的正态性假定是否成立。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，所用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直方图或正态概率图。**第(4)点即残差分析**

11.14 ★★★ 什么是置信区间估计和预测区间估计？两者有何区别？

12.简述线性回归中，置信区间和预测区间的区别？

01 利用估计的回归方程，给定一个 Xo，求出 y 的平均值的估计区间，则称为置信区间；若给定一个 Xo，求出 y 的某一个别值的估计区间，称为预测区间。

02 两个区间宽度不一，y 的平均值的预测区间要比平均值的预测区间更窄一些，这意味着，估计 y 的平均值要比预测 y 的某一个别值更准确。其次，可以发现，当 Xo=Xbar 时，估计就越准确，而 Xo 偏离 Xbar 越远，置信区间则越宽，估计的准确度就越差。

（计算题时，如何确定个别值与平均值区间估计？应看题目问的是置信区间还是预测区间，因为对参数（平均值）构造的是置信区间，对随机变量（个别值）构造的是预测区间。不必背诵，便于理解）

（缺）11.15 ★★★ 影响预测精度的因素有哪些？

个人认为，从公式$(\hat{y}_0-\delta,\hat{y}_0+\delta)$，$\delta=t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}$来看，影响一元线性回归预测精度的因素有：

1. 样本量大小
2. 总体的离散程度
3. 需要预测的自变量与样本自变量平均值的差异程度
4. 样本自变量的选取

（缺）11.16 ★★★ 回归分析的一般过程？

11.17 ★★★ 简述残差分析的作用

9 残差分析：残差分析就是通过残差所提供的信息，分析出数据的可靠性、周期性或其它干扰。用于分析模型的假定正确与否的方法。所谓残差是指观测值与预测值（拟合值）之间的差，即是实际观察值与回归估计值的差。

在回归分析中，测定值与按回归方程预测的值之差，以δ表示。残差δ遵从正态分布$N(0,\sigma^2)$。（δ-残差的均值）/残差的标准差，称为标准化残差，以δ^*表示。δ^*遵从标准正态分布$N(0,1)$。

在实际问题中，由于观察人员的粗心或偶然因素的干扰。常会使我们所得到的数据不完全可靠，即出现异常数据。有时即使通过相关系数或 F 检验证实回归方程可靠，也不能排除数据存在上述问题。残差分析的目的就在于解决这一问题。

11.18 ★★★ 相关分析和回归分析的联系和区别

10.相关分析与回归分析联系和区别

相关分析定义：用于判断两个或两个以上的变量之间是否存在相关关系，以及变量之间的关系形态如何，样本所反映的关系能否代表总体之间的关系的一种分析方法。

两者联系

01 二者有共同的研究对象，均是研究数值型因变量和数值型因变量之间的关系，二者可以相互补充。

02 相关分析可以表明变量之间的相关关系性质和程度，只有当变量之间存在着一定程度的相关关系时，进行回归分析去寻求变量之间的具体数学形式才有实际意义。

03 进行相关分析时，如果要具体确定变量之间相关的具体数学形式，又要依赖于回归分析。

两者区别

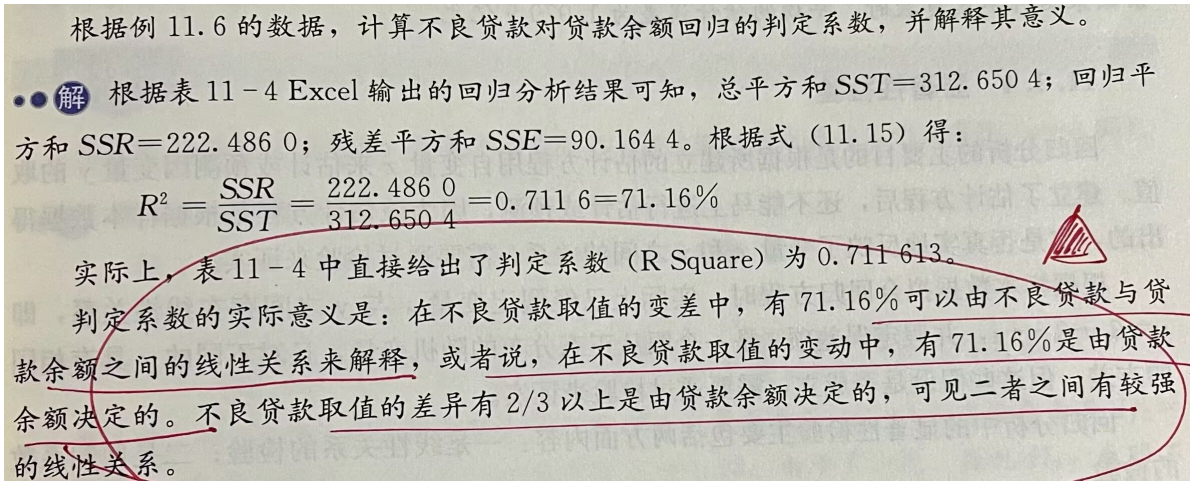
01 从研究目的上看，相关分析用一定的数量指标度量变量之间相互联系的方向和程度，回归分析则是寻求变量之间具体的数量关系式，并用一定的数学表达式与以表达。同时根据一个或者多个自变量的取值去预测或估计另一个特定变量的取值，并给出这种估计的可靠程度。

02 是从对变量处理上来看，相关分析对称的对待相关关系的变量，不考虑二者的因果关系，即不区分自变量和因变量，回归分析中不对称的对待相互联系的变量，即考虑二者因果关系，需要明确划分自变量和因变量。在相关分析中，两变量均视为随机变量，在回归分析中，通常假定自变量是取非固定的非随机变量。

11.19 为什么说不要用样本数据之外的x值预测y值?

因为在一元线性回归模型中，总是假定y与x之间的关系用线性模型来表达是正确的，但实际应用中，他们的关系可能是某种曲线。

补：判定系数的解释



根据例 11.6 的数据，计算不良贷款对贷款余额回归的判定系数，并解释其意义。

解 根据表 11－4 Excel 输出的回归分析结果可知，总平方和 $SST=312.650\ 4$；回归平方和 $SSR=222.486\ 0$；残差平方和 $SSE=90.164\ 4$。根据式（11.15）得：

$$R^2=\frac{SSR}{SST}=\frac{222.486\ 0}{312.650\ 4}=0.711\ 6=71.16\%$$

实际上，表 11－4 中直接给出了判定系数（R Square）为 0.711 613。

判定系数的实际意义是：在不良贷款取值的变差中，有 71.16%可以由不良贷款与贷款余额之间的线性关系来解释，或者说，在不良贷款取值的变动中，有 71.16%是由贷款余额决定的。不良贷款取值的差异有 2/3 以上是由贷款余额决定的，可见二者之间有较强的线性关系。

补：一元线性回归中判定系数与相关系数的联系与区别

(2015年重大432真题)

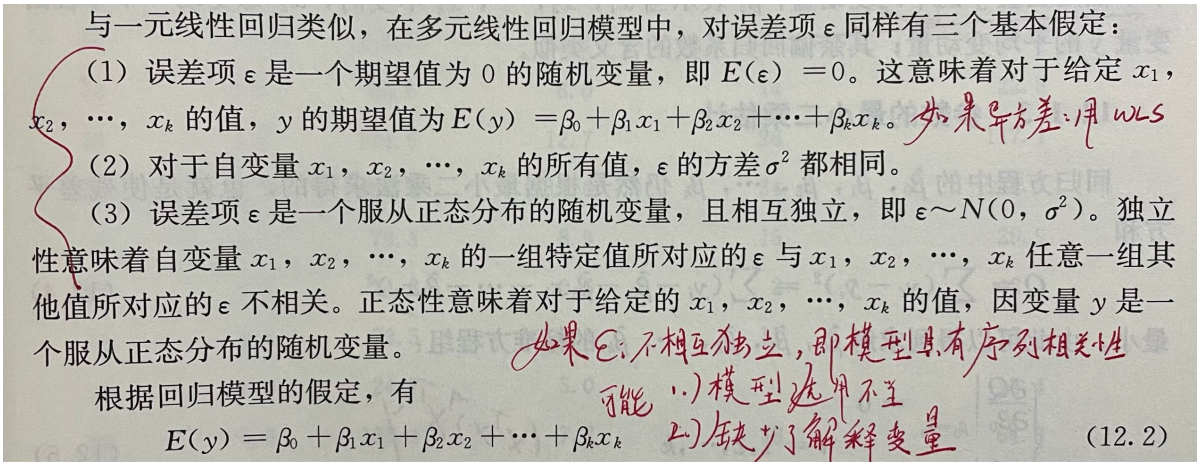
联系：在数值上，R^2等于r的平方

区别：

- 判定系数是就模型而言的，相关系数是针对两个变量而言的；
- 判定系数解释的是变量对变量的结实程度，相关系数则用于度量两个变量线性相关程度；
- 判定系数用于度量不对称的因果关系，相关系数则用于度量不含因果关系的对称相关关系；
- 判定系数的取值范围是$[0,1]$，相关系数的取值范围则是$[-1,1]$

Ch12 多元线性回归

12.1 ★★★ 多元线性回归模型的基本假定并简要说明假定不成立时如何应对



与一元线性回归类似，在多元线性回归模型中，对误差项 ε 同样有三个基本假定：

（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 x_1，x_2，…，x_k 的值，y 的期望值为 $E(y)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k$。如果异方差：用 WLS

（2）对于自变量 x_1，x_2，…，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，…，x_k 的一组特定值所对应的 ε 与 x_1，x_2，…，x_k 任意一组其他值所对应的 ε 不相关。正态性意味着对于给定的 x_1，x_2，…，x_k 的值，因变量 y 是一个服从正态分布的随机变量。

如果 ε_i 不相互独立，即模型具有序列相关性

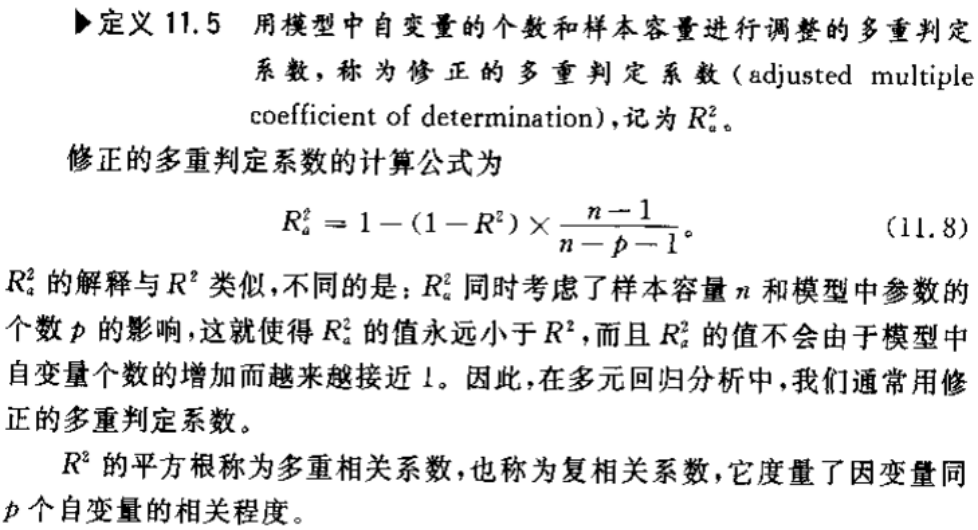
可能：1）模型选用不当

2）缺少了解释变量

根据回归模型的假定，有

$$E(y)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k \qquad (12.2)$$

12.2 ★★★ R^2和调整的R^2有何区别?



▶定义 11.5　用模型中自变量的个数和样本容量进行调整的多重判定系数，称为修正的多重判定系数（adjusted multiple coefficient of determination），记为 R_a^2。

修正的多重判定系数的计算公式为

$$R_a^2=1-(1-R^2)\times\frac{n-1}{n-p-1}。 \qquad (11.8)$$

R_a^2 的解释与 R^2 类似，不同的是：R_a^2 同时考虑了样本容量 n 和模型中参数的个数 p 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 1。因此，在多元回归分析中，我们通常用修正的多重判定系数。

R^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 p 个自变量的相关程度。

12.3 ★★★ 为什么要使用修正的判定系数?

1 什么是多重判定系数？为什么要计算和使用多重判定系数？

给出调整的 R^2 的计算公式为：$R_a^2 = 1-(1-R^2)\left(\frac{n-1}{n-k-1}\right)$

在多元回归分析中一般用调整 R^2 的来衡量模型的拟合效果，理由是：在多元回归分析中，自变量个数的增加将影响到因变量中被估计的回归方程所解释的变差数量。当增加自变量时，会使预测误差变得较小，从而减少残差平方和 SSE 。

由于回归平方和 $SSR = SST - SSE$ 。当 SSE 变小时，SSR 就会变大，从而使 R^2 变大。如果模型中增加一个自变量，即使这个自变量在统计上并不显著，R^2 也会变大。

因此，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 的解释与 R^2 类似，不同的是 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 1。

因此，在多元线性回归分析中，通常用调整的多重判定系数。

12.4 ★★★ 什么是多重共线性？它对回归分析有哪些影响?

2 多重共线性是什么？如何鉴别？如何修正？

多重共线性是指模型中两个或者两个以上的自变量彼此相关，称为多重共线性，当模型中存在多重共线性时，此时自变量会提供多余的信息，进而影响到模型的解释能力，可能使得回归结果分析混乱。主要表现在：

（1）多元线性回归模型中存在高度多重共线性产生的后果

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。

7. 多重共线性对参数估计的影响，以及对参数的假设检验和置信区间的影响。

多重共线性本身并不改变参数估计的无偏性性质，即平均估计值仍然等于真实参数值。然而，它会增加参数估计的不确定性，使得对参数的假设检验和参数区间估计变得更加困难。

参数假设检验：在存在多重共线性的情况下，参数的假设检验可能会受到影响。多重共线性使得参数估计的标准误差增加，导致检验统计量的值减小，从而增加了接受原假设（参数为零或无关）的可能性。这意味着在存在多重共线性的情况下，原本可能认为是显著的变量，在假设检验中可能变得不显著。因此，多重共线性可能会导致低估变量的重要性或显著性。

参数区间估计：多重共线性也会影响参数的置信区间估计。多重共线性导致参数估计的标准误差增加，进而使得置信区间变得更宽。这意味着在存在多重共线性的情况下，置信区间会更加容纳参数真实值的可能范围，估计值与真实值之间的偏差可能更大。

2018年重大432真题参考答案：多重共线性所产生的问题有

1. 自变量之间高度相关，这可能对回归的结果造成混乱，甚至把分析引入歧途
2. 多重共线性可能对参数估计值的正负号产生影响，特别是使回归系数β_i的正负号与预期相反

12.5 ★★★ 如何判别多重共线性?

（2）多重共线性常用的检验方法

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。具体来说，如果出现下列情况，暗示存在多重共线性：

a.模型中各对自变量之间显著相关。

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

c.回归系数的正负号与预期的相反。

d.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$1-R^2$。容忍度越小，多重共线性越严重。通常认为容忍度小于 0.1 时，存在严重的多重共线性。方差扩大因子等于容忍度的倒数，即$VIF=1/(1-R^2)$。显然，VIF越大，多重共线性越严重。一般认为VIF大于 10 时，存在严重的多重共线性。

2018年重大432真题参考答案：多重共线性的判别方法有

- 判断模型中各对自变量之间是否显著相关
- 是否有F检验显著而几乎对所有回归系数的t检验均不显著
- 回归系数的正负号是否与预期相反
- 计算VIF（方差膨胀因子），VIF越小，则多重共线性越严重

12.6 ★★★ 如何处理多重共线性?

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. 使用压缩估计，例如岭估计、Lasso估计等
3. 如果要在普通线性模型中保留所有的自变量，则应避免继续使用t检验，并且在对因变量y进行估计或预测时应该将自变量值限定在样本值的范围内

12.7 在多元线性回归中，选择自变量的方法有哪些?

- 向前选择
- 向后剔除
- 逐步回归
- 等等

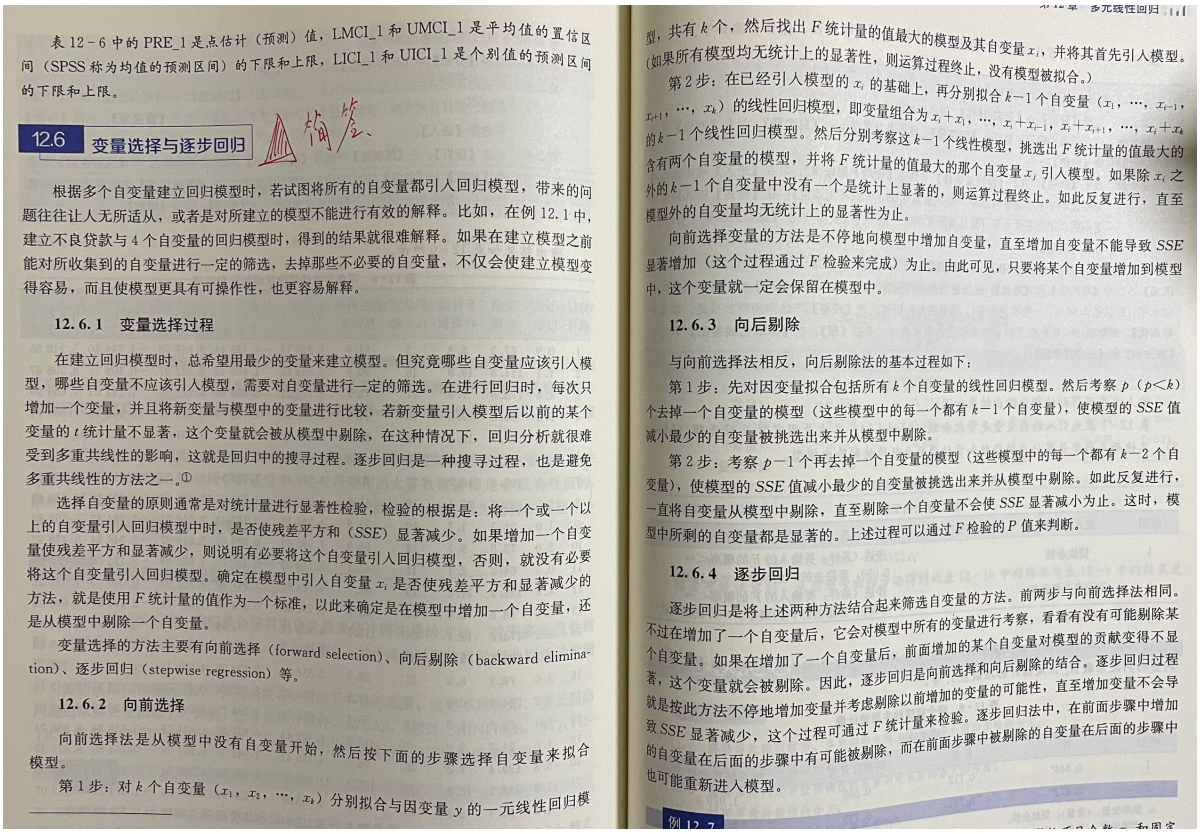


表 12－6 中的 PRE_1 是点估计（预测）值，LMCI_1 和 UMCI_1 是平均值的置信区间（SPSS 称为均值的预测区间）的下限和上限，LICI_1 和 UICI_1 是个别值的预测区间的下限和上限。

12.6 变量选择与逐步回归

根据多个自变量建立回归模型时，若试图将所有的自变量都引入回归模型，带来的问题往往让人无所适从，或者是对所建立的模型不能进行有效的解释。比如，在例 12.1 中，建立不良贷款与 4 个自变量的回归模型时，得到的结果就很难解释。如果在建立模型之前能对所收集到的自变量进行一定的筛选，去掉那些不必要的自变量，不仅会使建立模型变得容易，而且使模型更具有可操作性，也更容易解释。

12.6.1 变量选择过程

在建立回归模型时，总希望用最少的变量来建立模型。但究竟哪些自变量应该引入模型，哪些自变量不应该引入模型，需要对自变量进行一定的筛选。在进行回归时，每次只增加一个变量，并且将新变量与模型中的变量进行比较，若新变量引入模型后以前的某个变量的 t 统计量不显著，这个变量就会被从模型中剔除，在这种情况下，回归分析就很难受到多重共线性的影响，这就是回归中的搜寻过程。逐步回归是一种搜寻过程，也是避免多重共线性的方法之一。①

选择自变量的原则通常是对统计量进行显著性检验，检验的根据是：将一个或一个以上的自变量引入回归模型中时，是否使残差平方和（SSE）显著减少。如果增加一个自变量使残差平方和显著减少，则说明有必要将这个自变量引入回归模型，否则，就没有必要将这个自变量引入回归模型。确定在模型中引入自变量 x_i 是否使残差平方和显著减少的方法，就是使用 F 统计量的值作为一个标准，以此来确定是在模型中增加一个自变量，还是从模型中剔除一个自变量。

变量选择的方法主要有向前选择（forward selection）、向后剔除（backward elimination）、逐步回归（stepwise regression）等。

12.6.2 向前选择

向前选择法是从模型中没有自变量开始，然后按下面的步骤选择自变量来拟合模型。

第 1 步：对 k 个自变量（x_1，x_2，…，x_k）分别拟合与因变量 y 的一元线性回归模型，共有 k 个，然后找出 F 统计量的值最大的模型及其自变量 x_i，并将其首先引入模型。（如果所有模型均无统计上的显著性，则运算过程终止，没有模型被拟合。）

第 2 步：在已经引入模型的 x_i 的基础上，再分别拟合 $k-1$ 个自变量（x_1，…，x_{i-1}，x_{i+1}，…，x_k）的线性回归模型，即变量组合为 x_i+x_1，…，x_i+x_{i-1}，x_i+x_{i+1}，…，x_i+x_k 的 $k-1$ 个线性回归模型。然后分别考察这 $k-1$ 个线性模型，挑选出 F 统计量的值最大的含有两个自变量的模型，并将 F 统计量的值最大的那个自变量 x_j 引入模型。如果除 x_i 之外的 $k-1$ 个自变量中没有一个是统计上显著的，则运算过程终止。如此反复进行，直至模型外的自变量均无统计上的显著性为止。

向前选择变量的方法是不停地向模型中增加自变量，直至增加自变量不能导致 SSE 显著增加（这个过程通过 F 检验来完成）为止。由此可见，只要将某个自变量增加到模型中，这个变量就一定会保留在模型中。

12.6.3 向后剔除

与向前选择法相反，向后剔除法的基本过程如下：

第 1 步：先对因变量拟合包括所有 k 个自变量的线性回归模型。然后考察 p（$p<k$）个去掉一个自变量的模型（这些模型中的每一个都有 $k-1$ 个自变量），使模型的 SSE 值减小最少的自变量被挑选出来并从模型中剔除。

第 2 步：考察 $p-1$ 个再去掉一个自变量的模型（这些模型中的每一个都有 $k-2$ 个自变量），使模型的 SSE 值减小最少的自变量被挑选出来并从模型中剔除。如此反复进行，一直将自变量从模型中剔除，直至剔除一个自变量不会使 SSE 显著减小为止。这时，模型中所剩的自变量都是显著的。上述过程可以通过 F 检验的 P 值来判断。

12.6.4 逐步回归

逐步回归是将上述两种方法结合起来筛选自变量的方法。前两步与向前选择法相同。不过在增加了一个自变量后，它会对模型中所有的变量进行考察，看看有没有可能剔除某个自变量。如果在增加了一个自变量后，前面增加的某个自变量对模型的贡献变得不显著，这个变量就会被剔除。因此，逐步回归是向前选择和向后剔除的结合。逐步回归过程也可能重新进入模型。

例 12.7

补估计标准误差公式及其含义

多元回归中的标准误差：

$$S_e=\sqrt{\frac{SSE}{n-k-1}}=\sqrt{MSE} \tag{9}$$

标准误差S_e的含义是：当用所建立的多元线性回归方程做预测时，平均预测误差为$\sqrt{MSE}$个单位。

其实这就是σ的无偏估计量。

补一元线性回归公式速查

** 在有的教材中，对一元线性回归模型，规定：*

$$\begin{cases} l_{xx}=\sum(x_i-\bar{x})^2 \\ l_{yy}=\sum(y_i-\bar{y})^2 \\ l_{xy}=\sum(x_i-\bar{x})(y_i-\bar{y}) \end{cases} \tag{10}$$

于是

$$\hat{\beta}_1=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}=\frac{l_{xy}}{l_{xx}} \tag{11}$$

$$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x} \tag{12}$$

另外，在高斯-马尔可夫定理的条件下，有

$$(1)\quad \hat{\beta}_0 \sim N\left(\beta_0, \left(\frac{1}{n}+\frac{\bar{x}^2}{l_{xx}}\right)\sigma^2\right),\ \hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{l_{xx}}) \tag{13}$$

$$(2)\quad \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}}{l_{xx}}\sigma^2 \tag{14}$$

$$(3)\quad \hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \sim N\left(\beta_0+\beta_1 x_0, \left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}\right)\sigma^2\right) \tag{15}$$

$$(4)\quad \hat{\sigma}^2 = \frac{SSE}{n-2}\text{，这是无偏估计} \tag{16}$$

$$(5)\quad \mathbb{E}(SSR) = \sigma^2+\beta_1^2 l_{xx},\ \mathbb{E}(SSE) = (n-2)\sigma^2 \tag{17}$$

$$(6)\text{当}\beta_1=0\text{，有}\frac{SST}{\sigma^2}\sim\chi^2(n-1),\ \frac{SSR}{\sigma^2}\sim\chi^2(1),\ \frac{SSE}{\sigma^2}\sim\chi^2(n-2) \tag{18}$$

相应的，$SST=l_{yy}$，$SSR=\hat{\beta}_1^2 l_{xx}=\frac{l_{xy}^2}{l_{xx}}$，$SSE=SST-SSR$，在此一并给出参数显著性检验统计量：

$$F=\frac{(n-2)SSR}{SSE}=\frac{(n-2)l_{xy}^2}{l_{yy}l_{xx}-l_{xy}^2}\sim F(1,n-2) \tag{19}$$

$$t=\frac{\sqrt{SSR}}{\hat{\sigma}}=\sqrt{\frac{SSR}{\frac{SSE}{n-2}}}\sim t(n-2) \tag{20}$$

此外，对一元线性回归还有所谓相关系数检验，记$r=\frac{l_{xy}}{\sqrt{l_{xx}l_{yy}}}$为样本相关系数，置原假设为相关系数$\rho=0$，则

$$r^2\sim\frac{F(1,n-2)}{F(1,n-2)+n-2} \tag{21}$$

对一元线性回归而言，三个检验是等价的。

置信区间同理构造：

$$y_0\text{的置信区间：}(\hat{y}_0-\delta_0,\hat{y}_0+\delta_0),\quad \delta_0=t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}} \tag{22}$$

$$y_0\text{的预测区间：}(\hat{y}_0-\delta,\hat{y}_0+\delta),\quad \delta=t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}} \tag{23}$$

$$\beta_0\text{的置信区间：}(\hat{\beta}_0-\xi,\hat{\beta}_0+\xi),\quad \xi=t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{l_{xx}}} \tag{24}$$

$$\beta_1\text{的置信区间：}(\hat{\beta}_1-\eta,\hat{\beta}_1+\eta),\quad \eta=t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}\sqrt{\frac{1}{l_{xx}}} \tag{25}$$

补充一元线性回归证明

试试看以下几条现在还记得怎么证明吗？

- OLS的推导，即$\hat{\beta}_1=\frac{l_{xy}}{l_{xx}}$，$\hat{\beta}_0=\bar{y}-\bar{x}\hat{\beta}_1$（可以将导数为0式子列成线性方程，然后用克拉默法则求解）
- $\sum(y_i-\hat{y}_i)=\sum\varepsilon_i=0$，$\sum(y_i-\hat{y}_i)x_i=\sum\varepsilon_i x_i=0$
- $\hat{\beta}_1\sim N\left(\beta_1,\frac{\sigma^2}{l_{xx}}\right)$

- $\hat{\beta}_0 \sim N\left(\beta_0, \left(\frac{1}{n}+\frac{\bar{x}^2}{l_{xx}}\right)\sigma^2\right)$
- $\bar{y}$与$\hat{\beta}_1$相互独立
- $\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}}{l_{xx}}\sigma^2$
- $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \sim N\left(\beta_0+\beta_1 x, \left(\frac{1}{n}+\frac{(x-\bar{x})^2}{l_{xx}}\right)\sigma^2\right)$
- $SST = SSR + SSE$（注意用到第二条性质）
- $\frac{SSE}{n-2}$是σ^2的无偏估计（注意SSE的均值不方便直接求）
- 当$\beta_1 = 0$，有$\frac{SST}{\sigma^2} \sim \chi^2(n-1)$、$\frac{SSE}{\sigma^2} \sim \chi^2(n-2)$和$\frac{SSR}{\sigma^2} \sim \chi^2(1)$
- 假设检验统计量：$t = \frac{\sqrt{SSR}}{\hat{\sigma}} = \sqrt{\frac{SSR}{\frac{SSE}{n-2}}} \sim t(n-2)$，
  $F = \frac{(n-2)SSR}{SSE} = \frac{(n-2)l_{xy}^2}{l_{yy}l_{xx}-l_{xy}^2} \sim F(1, n-2)$，$r^2 \sim \frac{F(1,n-2)}{F(1,n-2)+n-2}$，三者等价

  前者利用$\frac{\hat{\beta}_1-\beta_1}{\sqrt{\frac{\delta^2}{l_{xx}}}} \sim N(0,1)$，中间者利用上一条性质，后者利用
  $SSE = SST - SSR = l_{yy}\left(1-\frac{l_{xy}^2}{l_{xx}l_{yy}}\right) = l_{yy}(1-r^2)$
- 额外补充：

$$y_0\text{的置信区间：}(\hat{y}_0-\delta_0, \hat{y}_0+\delta_0),\quad \delta_0 = t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}} \tag{26}$$

$$y_0\text{的预测区间：}(\hat{y}_0-\delta, \hat{y}_0+\delta),\quad \delta = t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}} \tag{27}$$

CH13 真题杂例 (时间序列与指数)

13.1 简述居民消费价格指数的作用

4.居民消费价格指数(CPI)：度量居民消费品和服务项目价格水平随时间变动的相对数，反映居民家庭购买的消费品和服务价格水平的变动情况。该指数是分析经济形势走势，检测物价水平，进行国民经济核算的重要指标，也常被用作测定通货膨胀。除此之外，还具有以下几方面作用：

①反映通货膨胀状况。通货膨胀的严重程度是用通货膨胀率来反映的，它说明了一定时期内商品价格持续上升的幅度。通货膨胀一般以居民消费价格指数来表示。计算公式为：

$$\text{通货膨胀率} = \frac{\text{报告期消费价格指数} - \text{基期消费价格指数}}{\text{基期消费价格指数}}$$

②反映居民购买力水平。货币购买力是指单位货币购买到的消费品和服务的数量。居民消费价格指数上涨，货币购买力则下降，反之则上升，货币购买力计算公式为：

$$\text{货币购买指数} = \frac{1}{\text{消费价格指数}} \times 100\%$$

③测定职工实际工资水平。居民消费价格指数的提高意味着实际工资的减少，居民消费价格指数下降意味着实际工资的提高。因此，利用居民消费价格指数可以将名义工资转化为实际工资。计算公式为：

$$\text{实际工资} = \frac{\text{名义工资}}{\text{消费价格指数}}$$

13.2 什么是零售价格指数、居民消费价格指数、生产价格指数、股票价格指数?

- 零售价格指数：指商品在市场上售卖价格变动的一种相对数
- 居民消费价格指数：指居民所购买的生活消费品和服务项目变化的一种相对数
- 生产价格指数：指商品在初级市场上出售的价格
- 股票价格指数：反应一定时期内，股票价格变动的一种相对数

13.3 一元线性回归的估计标准误差?

$$\sqrt{\frac{(1-r^2)l_{xy}}{n-2}} \tag{28}$$