概率论与数理统计

2020 年 5 月 24 日

目录

Chapter 1

导论

1.1 统计学及其应用领域

1.1.1 什么是统计学

统计学（statistics）是收集、处理、分析、解释数据并从数据中得出结论的科学
It deals with the collection, classification, analysis, and interpretation of numerical facts or data.
数据收集也就是取得数据
数据处理是将数据用图标等形式展示出来
数据分析是选择适当的统计方法研究数据，并从数据中提取有用信息进而得出结论
数据分析所用的方法可以分为描述性统计方法和推断统计方法
描述统计(descriptive statistics)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。
推断统计(inferential statistics)是研究如何利用样本数据来推断总体特征的统计方法。

1.1.2 统计的应用领域

1. 企业发展战略
2. 产品质量管理
3. 市场研究
4. 财务分析
5. 经济预测
6. 人力资源管理

统计更重要的功能是对数据进行分析
数据分析不是去寻找支持，不是先有结论再去寻找数据来支持结论
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
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

1.2 统计数据的类型

1.2.1 按照计量尺度的不同

分类数据(categorical data)：只能归于某一类别，非数字型数据
顺序数据(rank data)：只能归于某一有序类别，非数字
数值型数据(metric data)：按数字测量的观察值，表现为具体的数值
分类数据和顺序数据说明的是事物的本质特征，可以统称为定性数据或品质数据(qualitative data)
数值型数据说明的是现象的数量特征，因此也可称为定量数据或数量数据(quantitative data)

1.2.2 按照统计数据的收集方法

观测数据(observational data)：通过调查或观测收集到的数据
实验数据(experimental data)：在实验中控制实验对象而收集到的数据。

1.2.3 按照被描述现象与时间的关系

截面数据(cross-sectional data)：在相同或近似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
时间序列数据(time series data)：在不同时间点上收集的数据，用于描述现象随时间变化的情况

1.2.4 区分数据类型的作用

不同的数据需要用不同的统计方法来处理和分析
分类数据：计算各组频数或频率，计算众数和异众比率，进行列联表分析和 χ^2 检验等等
顺序数据：计算中位数和四分位差，计算等级相关系数等等
数值型数据：计算各种统计量，进行参数估计和检验等等

1.3 统计学中的几个基本概念

1.3.1 总体和样本

总体(population)是包含所研究的全部个体（数据）的集合
总体可分为有限总体和无限总体，目的是判别在抽样中每次抽取是否独立。
样本(sample)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(sample size)

1.3.2 参数和统计量

参数(parameter)是用来描述总体特征的概括性数字度量，他是研究者想要了解的总体的某种特征值。通常用希腊字母表示，参数是一个未知的常数。

统计量(statistic)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（抽样是随机的），样本统计量通常用英文字母表示。统计量是已经知道的（样本已经抽取出来了）

常用的统计量有样本均值、样本比例、样本方差、z统计量、t统计量、χ^2统计量、F统计量等等
抽样的目的就是要根据样本统计量去估计总体参数。

1.4 变量

变量(variable)是说明现象某种特征的概念，特点是从一次观测到下一次观测结果会呈现出变化

1.4.1 变量的分类

分类变量(categorical variable)是说明事物类别的一个名称，取值是分类数据

顺序变量(rank variable)是说明事物有序类别的一个名词，其取值是顺序数据，比如说一个人对某种事物的看法

数值型变量(metric variable)是说明事物数字特征的一个名称，其取值是数值型数据。

随机变量和非随机变量

经验变量(empirical variable)描述周围环境中可以观察到的事物

理论变量(theoretical variable)是由统计学家用数学方法构造出来的一些变量，比如z统计量、t统计量、χ^2统计量、F统计量等等

Chapter 2

数据的搜集

2.1 数据的来源

2.1.1 数据的间接来源

与研究内容有关的原信息已经存在，只需要对其进行加工、整理，则称之为间接来源的数据
优点：搜集容易、采集成本低、很快得到
局限性：

2.1.2 数据的直接来源

通过调查（调查数据）和实验（实验数据）获得的数据
优点：针对性强、直接

2.2 调查方法

2.2.1 概率抽样和非概率抽样

进行什么样的抽样首先取决于研究目的。一个好的样本应该具有最好的性能价格比。

概率抽样

概率抽样（probability sampling）也称为简单抽样，是指遵循随机原则进行的抽样，总体中每个样本都有一定机会被选入样本，特点：

1. 抽样时按照一定的概率以随机原则抽取样本
2. 每个单位被抽中的概率时已知的，或者可以计算出来
3. 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

优点　1. 可以依据调查结果计算估计误差，从而得到对总体目标量进行推断的可靠程度

简单随机抽样　（simple random sampling）就是从包括总体N个单位的抽样框中随机地、一个个的抽取n个单位作为样本，每个单位的入样概率是相等的。

分层抽样　（stratified sampling）是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。

整群抽样 （cluster sampling）将总体中若干个单位合并为组，这样的组称为群。抽样时直接抽取群，然后对中选群中所有单位实施调查。

整群抽样与分层抽样的区别 分层抽样是按照相似的合并为层，整群抽样可以不按照相似来分类

分层抽样分层后要从不同的层中独立、随机的抽取样本

整群抽样分群后直接抽样群，然后对中选群中所有单位实施调查

系统抽样 (systematic sampling)将总体中的所有单位（抽样单位）按照一定顺序排列，在规定的范围内随机抽取一个单位作为初始单位，然后按照事先制定好的规则确定其他样本单位。

多阶段抽样 (multi-stage sampling)类似整群抽样，首先抽取群，然后进一步抽样，从选中的群中抽取若干个单位进行调查（二阶段抽样），阶段增多，就称为多阶段抽样。

概率抽样最主要的优点：可以根据调查结果计算调查误差，从而得到对总体目标量进行推断的可靠程度。

非概率抽样

非概率抽样(non-probability sampling)指抽取时不是依据随机抽样原则，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。

方便抽样 调查中调查员依据方便的原则，自行确定作为样本的单位。

判断抽样 研究人员依据经验、判断和对研究对象的了解，有目的的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、典型抽样、代表抽样等方式。

自愿样本 被调查者自愿参加

滚雪球抽样 往往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。

配额抽样 类似与分层抽样，首先将总体中的所有单位按一定的标志（变量）分为若干类，然后在每个类中采用方便抽样或判断抽样的方式选取样本单位。

概率抽样与非概率抽样的比较

1. 概率抽样是依据随机原则抽取样本

   概率抽样的样本统计量的理论分布存在，可以根据调查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

   技术含量较高、调查成本高

2. 非概率抽样不是依据随机原则抽取样本

   所以非概率抽样的样本统计量的分布不明确，无法使用样本结果对总体相应的参数进行推断

   操作简单、时效快、成本低、对于抽样中的统计要求不是很高

2.2.2 收集数据的基本方法

自填式 没有调查员协助的情况下由被调查者自己填写，完成调查问卷。

面访式 现场与调查员面对面，调查员提问、被调查者回答

电话式 调查人员通过电话向被调查者实施调查。

数据收集方法的选择：

1. 抽样框中的有关信息
2. 目标总体的特征
3. 调查问卷的内容
4. 有形辅助物的使用
5. 实施调查的自愿
6. 管理与控制
7. 质量要求

补充一个表格，p23

2.3 实验方法

2.3.1 实验组和对照组

实验法的基本逻辑：有意识的改变某个变量A的情况，然后看另一个变量B变化的情况。如果B会随着A的变化而变化，那么就说明A对B有影响。

实验组(experiment)是指随机抽选的对象的子集

对照组(control group)每个单位不接受实验组所接受的某种特别处理

2.3.2 实验中的若干问题

1. 人的意愿
2. 心理问题
3. 道德问题

2.3.3 实验中的统计

2.3.4 实验法案例

2.4 数据的误差

2.4.1 抽样误差

抽样误差(sampling error)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。

2.4.2 非抽样误差

抽样框 用于抽选样本的总体单位信息，是概率抽样中必不可少的

抽样框误差 由于抽样框的不完善而造成的误差

回答误差 由于被调查者在接受调查时的回答与真实值不符合造成的误差

无回答误差 被调查者拒绝接受回答造成的误差

调查员误差 由于调查员的原因而产生的误差

测量误差 与测量工具有关的误差

2.4.3 误差的控制

2.4.4 做习题时想到要补充的概念

二手数据的特点

1. 收集容易、采集数据成本低、能很快得到
2. 分析所要研究的问题
3. 提供研究的背景
4. 帮助研究者更好的定义问题，检验和回答某些假设和疑问
5. 寻找研究问题的思路和途径

抽样框 (sampling frame)

1. 通常包含所有总体单位的信息
2. 提供备选单位以供抽选
3. 计算各个单位入样概率的依据

简单随机抽样 （simple random sampling）就是从包括总体N个单位的抽样框中随机地、一个个的抽取n个单位作为样本，每个单位的入样概率时相等的。

分层抽样 （stratified sampling）是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。

配额抽样 类似与分层抽样，首先将总体中的所有单位按一定的标志（变量）分为若干类，然后在每个类中采用方便抽样或判断抽烟的方式选取样本单位。

系统抽样 (systematic sampling)将总体中的所有单位（抽样单位）按照一定顺序排列，在规定的范围内随机抽取一个单位作为初始单位，然后按照事先制定好的规则确定其他样本单位。

整群抽样 （cluster sampling）将总体中若干个单位合并为组，这样的组称为群。抽样时直接抽取群，然后对中选群中所有单位实施调查。

多阶段抽样 (multi-stage sampling)类似整群抽样，首先抽取群，然后进一步抽样，从选中的群中抽取若干个单位进行调查（二阶段抽样），阶段增多，就称为多阶段抽样。

判断抽样 研究人员依据经验、判断和对研究对象的了解，有目的的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、典型抽样、代表抽样等方式。

重复抽样 没有这个概念

不重复抽样 没有这个概念

自愿抽样 没有自愿抽样，只有自愿样本

自愿样本 被调查者自愿参加，成为样本中的一分子，向调查人员提供有关信息

方便抽样 调查中调查员依据方便的原则，自行确定作为样本的单位。

滚雪球抽样 往往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。

概率抽样 也称随机抽样，是指遵循随即原则进行的概率抽样，总体中每一个单位都有一定的机会被选入样本。

特点
1. 抽样时按一定的概率以随机原则抽取样本
2. 每个单位被抽中的概率是已知的，或者可以计算出来
3. 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

分类
1. 简单随机抽样
2. 分层抽样
3. 整群抽样
4. 系统抽样
5. 多阶段抽样

非概率抽样 非概率抽样(non-probability sampling)指抽取时不是依据随机抽样原则，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。

特点
1. 操作简单、时效快、成本低
2. 对于抽样中的统计专业技术要求不是很高
3. 不是依据随即原则抽取样本
4. 样本统计量的分布是不确切的
5. 无法使用样本的结果对总体相应的参数进行推断

分类
1. 方便抽样
2. 判断抽样
3. 自愿样本
4. 滚雪球抽样
5. 配额抽样

概率抽样调查

观察调查

实验调查

Chapter 3

数据的图表表示

3.1 数据的预处理

3.1.1 数据审核

检查数据中是否有错误
通过调查取得的原始数据(raw data)要从完整性和准确性两个方面去审核

1. 完整性审核
2. 准确性审核

通过其他渠道取得的二手数据，应该着重审核数据的适用性和时效性

1. 适用性
2. 时效性

3.1.2 数据筛选

(data filter)
根据需要找出符合条件的某类数据

3.1.3 数据排序

3.1.4 数据透视表

3.2 品质数据的整理与展示

品质型数据主要是做分类整理

3.2.1 分类数据的整理与图示

频数与频数分布

分类数据的图示

1. 条形图
2. 帕累托图：按照各类别数据出现的频数多少排序后绘制的条形图
3. 饼图
4. 环形图

3.2.2 顺序数据的整理与图示

累积频数 (cumulative)是将各有序类别或组的频数逐级累加起来得到的频数

累积频率 (cumulative percentages)是将各有序类别或组的百分比逐级累加起来

累积频数和累积频率类似于数理统计中的经验分布函数

与经验分布函数不同的一点是累积频数和累积频率有向上累积和向下累积两种方式

3.3 数值型数据的整理与展示

数值型数据主要是做分组整理

3.3.1 数据分组

单变量分组 每个变量作为一组，适合离散变量

组距分组 将全部变量依次划分区间，并将一个区间的变量值作为一组

3.3.2 数值型数据的图示

分组数据 直方图(histogram)

未分组数据

1. 茎叶图(stem-and-leaf display)：是反应**原始数据分布**
2. 箱线图(box plot)：反应**原始数据分布的特征**

茎叶图，将一个数字分为两个部分

箱线图，根据一组数据的最大值、最小值、中位数、两个四分位数这五个特征值绘制而成的

时间序列数据 线图(line plot)，主要用于反应现象随时间变化的特征

多变量数据

1. 散点图(scatter disgram)，用二维坐标展示两个变量关系
2. 气泡图(bubble chart)，用于展示三个变量之间的关系(气泡大小是第三个维度)
3. 雷达图(rader chart)，也成为蜘蛛图(spyder chart)，显示多个变量

3.4 总结：数据的类型与主要图示方法

数据类型： 品质数据、数值型数据

品质数据： 汇总表

汇总表： 条形图、饼图、环形图

数值型数据： 原始数据、分组数据、时间序列数据、多变量数据

原始数据： 茎叶图、箱线图

分组数据： 直方图

时间序列数据： 线图

多变量数据： 散点图、气泡图、雷达图

3.5 合理使用图表

设计图形时，应该绘制得尽可能简洁，以清晰的显示数据、合理的表达统计目的为依据

3.5.1 鉴别图像优劣的准则

3.5.2 统计表的设计

Chapter 4

数据的概括性度量

数据的度量	分类数据	顺序数据	数值型数据
集中趋势的度量	众数	中位数、分位数	平均数
离散程度的度量	异众比率	四分位差	极差、平均差、方差、标准差
分布形状的测度	偏态、峰态		
相对位置的度量	标准分数、经验法则、切比雪夫不等式		
相对离散程度的度量	离散系数		

4.1 k阶原点矩与k阶中心矩

k阶原点矩 $\mu_k = E(X^K)$

k阶中心矩 $v_k = E(X - E(X))^k$

这一个总结是根据贾俊平的第七版统计学来总结的，在第四章里面所有的计算公式都是按照离散型随机变量来计算的

看茆诗松的书里面有连续性随机变量的计算公式

4.2 集中趋势的度量

4.2.1 分类数据：众数

众数 (mode)是一组数据中出现次数最多的变量值，用M_0表示

众数主要用于测度分类数据的集中趋势

众数是一个位置代表值，不受极端值影响，众数有可能有多个

4.2.2 顺序数据：中位数和分位数

中位数 (median)是一组数据排序后处于中间位置上的数据，用M_e表示

中位数主要用于测度顺序数据的集中趋势，也适合测度数值型数据的集中趋势

$$M_e = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & odd \\ \frac{1}{2}\left\{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right\} & even \end{cases}$$

四分位数 (quartitle)，是一组数据排序后处于25%和75%位置上的值

Q_L位置= $\frac{n}{4}$

Q_U位置= $\frac{3n}{4}$

4.2.3 数值型数据：平均数

平均数 也称均值(mean)，是集中趋势的最主要测度值

简单平均数 未经分组数据计算的平均数

$$\overline{x} = \frac{x_1 + x_2 + \cdot + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

加权平均数 分组数据计算的平均数，M_i表示各组的组中值，f_i表示频数

$$\overline{x} = \frac{M_1 f_1 + M_2 f_2 + \cdot + M_k f_k}{f_1 + f_2 + \cdot + f_k} = \frac{\sum_{i=1}^{k} M_i f_i}{n}$$

几何平均数 (geometric mean) 是n个变量值乘积的n次方根，用G表示

$$G = \sqrt[n]{x_1 \times x_2 \times ... \times x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

几何平均数主要用于计算平均比率
在实际应用中，几何平均数主要用于计算现象的平均增长率
当变量值本身是比率形式时，采用几何平均要比算数平均更合理

4.2.4 众数、中位数和平均数的比较

关系 1. 对称分布 $M_0 = M_e = \overline{x}$

2. 左偏分布 $\overline{x} < M_e < M_0$

3. 右偏分布 $M_0 < M_e < \overline{x}$

这里应该插入教材76页的图片

特点 1. 众数不受极端值的影响，不具有唯一性，在数据较多的时候才有意义

2. 中位数不受极端值的影响

3. 平均数易受极端值的影响，对于偏态分布其代表性较差

4.3 离散程度的度量

4.3.1 分类数据：异众比率

异众比率 (variation ratio) 是指非众数组的频数占总频数的比例，用V_r表示

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

$\sum f_i$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

特点 1. 异众比率大，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

2. 异众比率小，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

4.3.2 顺序数据：四分位差

四分位差 (quartile deciation)也成为内距或四分间距，是上四分位数与下四分位数只差，用Q_d表示

$$Q_d = Q_U - Q_L$$

四分位差反映了中间50% 的数据的离散程度

越小越集中

不适合分类数据

4.3.3 数值型数据：极差、平均差、方差和标准差

极差 (range)，一组数据最大值与最小值之差，用R表示

$$R = \max x_i - \min x_i$$

平均差 (mean deviation)，也称平均绝对离差，用M_d表示

1. 未分组数据 $M_d = \frac{\sum_{i=1}^{n}|x_i - \overline{x}|}{n}$
2. 分组数据 $M_d = \frac{\sum_{i=1}^{k}|M_i - \overline{x}|f_i}{n}$

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程度，能全面准确的反映一组数据的离散状况

平均差越大越离散

方差 (variance) 是各变量值与其平均数离差平方和的平均数

样本方差s^2

1. 未分组数据 $s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$
2. 分组数据 $s^2 = \frac{\sum_{i=1}^{k}(M_i - \overline{x})^2 f_i}{n-1}$

方差能较好的反映出数据的离散程度

标准差 是有量纲的，与变量值的计量单位相同

1. 未分组数据 $s^2 = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$
2. 分组数据 $s^2 = \sqrt{\frac{\sum_{i=1}^{k}(M_i - \overline{x})^2 f_i}{n-1}}$

标准差是的作用是统一量纲

相对位置的度量

1. 标准分数(standard score) $z_i = \frac{x_i - \overline{x}}{s}$
2. 经验法则（提供范围）
3. 切比雪夫不等式（提供下界，至少）

标准分数 (standard score)

$$z_i = \frac{x_i - \overline{x}}{s}$$

类似正态分布的标准化中心化

给出一组数据中各数据的相对位置

经验法则 当一组数据对称分布时，经验法则表明：

3σ法则
1. 约有68%的数据在平均数±1个标准差的范围内
2. 约有95%的数据在平均数±2个标准差的范围内
3. 约有99%的数据在平均数±3个标准差的范围内

切比雪夫不等式 (Chebyshev's inequality)

$$p\{|X-EX|<\varepsilon\} \quad \geq \quad 1-\frac{VarX}{\varepsilon^2}$$
$$p\{|X-EX|<k\sigma\} \quad \geq \quad 1-\frac{1}{k^2}$$

对于任意分布形态的数据，根据切比雪夫不等式，至少有$(1-1/k^2)$的数据落在±k个标准差之内，k为大于1的任意值

对于k=2,3,4，该不等式的含义是

切比雪夫
1. 至少有75%的数据在平均数±2个标准差的范围内 $p\{|X-EX|<2\sigma\} \geq 1-\frac{1}{2^2}=\frac{3}{4}$
2. 至少有89%的数据在平均数±3个标准差的范围内 $p\{|X-EX|<3\sigma\} \geq 1-\frac{1}{3^2}=\frac{8}{9}$
3. 至少有94%的数据在平均数±4个标准差的范围内 $p\{|X-EX|<4\sigma\} \geq 1-\frac{1}{4^2}=\frac{15}{16}$

4.3.4 相对离散程度：离散系数

离散系数 (coefficient of variation)，也称为变异系数，记为V_s

$$V_s=\frac{s}{\overline{x}}$$

离散系数用于比较不同样本数据的离散程度

采用不同单位计计量的变量值，其离散程度的测度值也就不同，因此，对于平均水平不同或计量单位不同的不同组别的变量值，不能用标准差直接比较其离散程度

离散系数越大，数据的离散程度越大

根据这个公式，**两组数据若标准差相等，平均数小的离散程度大**

4.4 偏态与峰态的度量

1. k阶原点矩 $\mu_k=E(X^K)$
2. k阶中心矩 $v_k=E(X-E(X))^k$

4.4.1 偏态及其测度

偏态 (skewness)是对数据分布对称性的测度

偏态系数 (coefficient of skewness)是测度偏态的统计量，记作SK

1. 未分组数据 $SK=\frac{n\sum(x_i-\overline{x}^3)}{(n-1)(n-2)s^3}$
2. 分组数据 $SK=\frac{\sum_{i=1}^{k}(M_i-\overline{x})^3f_i}{ns^3}$

$$\beta_s=\frac{v_3}{v_2^{3/2}}$$

数据对称，$SK=0$

数据右偏，$SK>0$

数据左偏，$SK<0$

分布对称时，离差的三次方正负抵消，所以为零

SK的数值越大，偏斜的程度越大

4.4.2 峰态及其测度

峰态 (kurtosis)是对数据分布平峰或尖峰程度的测度

峰态系数 (coefficient of kurtosis)是测度峰态系数的统计量，记作K

$$\beta_k = \frac{v_4}{v_2^2}$$

峰态通常是与标准正态分布相比较而言的

标准正态分布的峰态系数等于3（若原来的公式中没有减去3）

$K > 3$时为尖峰分布，数据分布更集中

$K < 3$时为扁平分布，数据分布越分散

4.5 总结

数据的度量	分类数据	顺序数据	数值型数据
集中趋势的度量	众数	中位数、分位数	平均数
离散程度的度量	异众比率	四分位差	极差、平均差、方差、标准差
分布形状的测度	偏态、峰态		

相对位置的度量	标准分数、经验法则、切比雪夫不等式
相对离散程度的度量	离散系数

Chapter 5

概率与概率分布

5.1 随机事件及其概率

推断统计 在搜集、整理和观测样本数据的基础上，对有关总体做出推断

特点是根据随机的观测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

5.1.1 随机事件的几个基本概念

试验 对某事物或现象进行的观察或实验

事件 观察或试验的结果

随机事件 (random event) 在同一组条件下，每次试验有可能出现也有可能不出现的事件

必然事件

不可能事件

基本事件 (elementary event)一个事件不能被分解为两个或者更多的事件

5.1.2 事件的概率

事件出现的可能性大小叫做概率

1. **概率的古典定义** $P(A)=\frac{\text{事件A所包含的基本事件个数}}{\text{样本空间所包含的基本事件个数}}=\frac{m}{n}$
2. **概率的统计定义** $P(A)=\frac{\text{事件A出现m次}}{\text{随机试验n次}}=\frac{m}{n}=p$
3. **主观概率定义** 对一些无法重复的实验，只能根据以往经验人为确定这个事件的概率

5.1.3 随机变量的概念

在同一组条件下，如果每次实验可能出现这样或者那样的结果，并且所有的结果都能列举出来，即X的所有可能值 $x_1,x_2,...,x_n$都能列举出来，而且X的可能值$x_1,x_2,...,x_n$具有确定概率 $P(x_1),P(x_2),...,P(x_n)$，其中$P(x_i)=P(X=x_i)$，称为概率函数(probability function), 则X称为$P(X)$的随机变量， $P(X)$称为随机变量X的概率函数

随机变量的概念是由概率函数连在一起的

首先定义概率函数

概率函数中的自变量就是随机变量

知乎定义：随机变量（random variable）表示随机现象（在一定条件下，并不总是出现相同结果的现象称为随机现象）中各种结果的实值函数（一切可能的样本点）。

知乎问题：如果用通俗的语言介绍什么什么是随机变量？

通俗地讲，随机变量就是一个随机的数，它是对任何的“随机的东西”做的量化。我相信你有能力解释什么是“随机”，所以主要解释“量化”的部分。

随机的对象可以是任何东西–明天的天气可以是晴、阴、雨，扔硬币的结果可以是正面或者反面，这里本身都没有数字。但是我们要借助概率论来研究它们，而概率论是数学的一部分，要用到数学语言，那么总是写“明天是晴天的概率”就很不方便，于是我们可以把晴、阴、雨贴上标签，叫做0、1、2，然后把明天的天气状况用一个字母X来表示，于是“明天下雨”就变成了“X=2”。这样，这个原本没有数字的随机结果就变成了一个可能的取值为0、1、2的随机数，这就是随机变量。

作者：Yves S

链接：https://www.zhihu.com/question/307188808/answer/566226225

来源：知乎

参考这个专栏：【测度论】概率论与测度论之间联系的通俗解释(一) https://zhuanlan.zhihu.com/p/23629928

不是每个事件都可以定义其概率（发生的可能性的大小）的，对应的就是不是每个集合都可以定义测度，可以定义测度集合就是可测集。同时，事件必然要涉及到事件的组合运算（复杂事件是可由基本事件表示出来），对应的就是集合的交、并、差、余、极限的运算到复杂集合，所以又需要保证做可列次这些运算不能超出全体范围（即可测集的范围要足够大，以保证集合的可列次交、并、差、余、极限的运算，之后还在里面）

随机变量的测度论语言定义：设$(X,\mathcal{F},P)$为概率测度空间，若对实数轴上 Borel σ-代数中的任一集合（(Borel集)B，都有$\{\omega\in\Omega:X(\omega)\in B\}\in\mathcal{F}$，则称$X(\omega)$为随机变量，简记为$X$

随机变量就是建立了“随机事件”到“实数轴上Borel σ-代数”的一种对应，并且保证了建立了这种对应的随机事件都是可以定义概率测度的。

1. 离散型随机变量
2. 连续性随机变量

5.2 离散型随机变量及其分布

随机变量X的所有取值都可以逐个列举出来

1. 期望 $\mu=x_1p_1+x_2p_2+...+x_np_n=\sum_{i=1}^{n}x_ip_i$
2. 方差 $\sigma^2=D(X)=E[(X-E(X))]^2=\sum_{i=1}^{\infty}[x_i-E(X)]^2p_i$
3. 离散系数 $V=\frac{\sigma}{\mu}$

离散系数可用来比较不同期望值的总体之间的离中趋势

方差的一个简化公式

$$\sigma^2=D(X)=E(X^2)-[E(X)]^2$$

方差实际上就是随机变量X的函数$[X-E(X)]^2]$的数学期望

5.2.1 二项分布

(binomial distribution)

假设 1. 包含n个相同的试验

2. 每次实验只有两个可能的结果
3. 出现“成功”的概率p对每一次试验是相同的
4. 试验相互独立试验“成功”、“失败”可以计数

n重伯努利试验 具有上述特征的n次独立重复试验

X表示n次重复独立试验中事件A（成功）出现的次数，记作$X \sim B(n,p)$

$$P\{X=k\} = C_n^k p^k q^{n-k} = \frac{n!}{k!(n-k)!} p^k p^{n-k}$$

$$E(X) = np$$

$$D(X) = npq$$

5.2.2 泊松分布

(Poisson distribution) 用来描述在一指定时间范围内或在指定面积内某一事件出现的次数的分布

$$P(X=k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

$$E(X) = \lambda$$

$$D(X) = \lambda$$

5.2.3 二项分布的泊松近似

在n重伯努利试验中，当成功的概率很小$(p \to 0)$，试验次数很大时，二项分布近似等于泊松分布，即

$$C_N^k p^k q^{n-k} \approx \frac{\lambda^k}{k!} e^{-\lambda}$$

实际应用中，当$p \le 0.25, n > 20, np \le 5$时，用泊松分布近似二项分布效果较好

5.3 连续性随机变量的概率分布

随机变量X的所有取值无法逐个列举出来

概率密度函数$f(X)$满足：

1. 非负性$f(x) \ge 0$
2. 正则性$\int_{-\infty}^{\infty} f(x)dx = 1$

分布函数$F(X)$满足：

1. 单调性 $F(x)$是定义在整个实数轴上的单调非减函数
2. 有界性

   $\lim\limits_{x \to -\infty} F(x) = 0$

   $\lim\limits_{x \to \infty} F(x) = 1$
3. 右连续性

分布函数$F(x)$定义：

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx \quad -\infty < x < \infty$$

$$P(a < X < b) = \int_{a}^{b} f(x)dx = F(b) - F(a)$$

$$f(x) = F^{'}(x)$$

期望 $\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$

方差 $\sigma^2 = D(X) = \int_{-\infty}^{\infty}[x - E(x)]^2 f(x)dx$

5.3.1 正态分布

(normal distribution)正态分布概率密度函数为：

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \quad -\infty < x < \infty$$

记作$X \sim N(\mu, \sigma^2)$

5.3.2 标准正态分布

(standard normal distribution)标准正态分布的概率密度函数为：

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad -\infty < x < \infty$$

$\psi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$
$\Phi(x) = \int_{-\infty}^{x} \varphi(t)dt = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$

5.3.3 标准正态分布的重要性

任何一个一般的正态分布都可以通过线性变换转化为标准正态分布
设$X \sim N(\mu, \sigma^2)$，则

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

对于负值，可以采取如下公式

$$\Phi(-x) = 1 - \Phi(x)$$

5.3.4 正态分布在质量管理中的应用

6σ准则：

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = \Phi(3) - \Phi(-3) = 0.9973$$

当某质量特性$X \sim N(\mu, \sigma^2)$时，其特性值落在区间 $(\mu - 3\sigma, \mu + 3\sigma)$ 外的概率仅为0.27%.

Chapter 6

统计量及其抽样分布

6.1 统计量

定义 设$X_1, X_2, ..., X_n$是从总体中抽取的容量为n的一个样本，如果由此样本构造的一个函数$T(X_1, X_2, ..., X_n)$，不依赖任何未知参数，则称函数$T(X_1, X_2, ..., X_n)$是一个统计量

特点
1. 统计量是样本的函数
2. 统计量不依赖于任何未知参数
3. 统计量是对原始数据进行一定的计算，得到具有代表性的数字，用来反映数据某些方面的特征
4. 总体参数，参数是描述总体特征的概括性数字度量
5. 样本统计量，统计量是描述样本特征的概括性数字度量
6. 统计量的意义是根据样本统计量去估计总体参数，或者描述总体
7. 统计量在统计学中具有极其重要的地位，是统计推断的基础
8. 统计量在统计学中的地位相当于随机变量在概率论中的地位

6.1.1 常用的统计量

样本均值 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

样本方差/标准差 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{x})^2$

样本变异系数 $V = S/\overline{X}$

样本k阶(原点)矩 $\mu_k = E(X^K)$

样本k阶中心矩 $v_k = E(X - E(X))^k$

偏度系数 coefficient of skewness $\beta_s = \frac{v_3}{v_2^{3/2}}$

峰度系数 coefficient of kurtosis $\beta_k = \frac{v_4}{v_2^2}$

6.1.2 用于检验的统计量

z统计量 (已知总体标准差) (小样本已知总体标准差)

$$z = \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

(大样本未知总体标准差)

$$z = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

(总体比例的检验)

$$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

(两个总体均值的检验，已知两个总体的方差)

$$z = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

(两个总体比例之差的检验，检验两个总体比例相等)

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

t统计量 (小样本，未知总体标准差)

$$t = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

(两个总体均值的检验，未知两个总体的方差但是可知两个总体方差相等，n较小)

$$z = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

自由度为$n_1 + n_2 - 2$

χ^2统计量 (总体方差的检验) (小样本已知总体标准差)

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

F统计量 (两个方差比)

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

分类数据的χ^2统计量

$$\chi^2 = \Sigma\frac{(f_0 - f_e)^2}{f_e}$$

6.2 由正态分布导出的几个重要分布

6.2.1 抽样分布

定义 在总体的分布类型已知时，若对任一自然数n，都能导出统计量$T = T(X_1, X_2, ..., X_n)$ 的分布的数学表达式，这种分布称为精确的抽样分布。

抽样分布就是抽样的样本的分布，可以理解为抽取出的样本服从的分布

精确的抽样分布在样本量较小的时候很有用

精确的抽样分布大多数是在正态总体情况下得到的

在正态总体下，主要有χ^2分布、t分布、F分布

⇓

6.2.2 χ^2分布

定义 1. 随机变量$X_1, X_2, ..., X_n$相互独立

2. X_i服从标准正态分布$N(0,1)$

则其平方和$\sum_{i=1}^{n}$服从自由度为n的χ^2分布

自由度解释为独立变量的个数，也可以解释为二次型的秩

χ^2分布的自由度n越大，其概率密度曲线趋于对称

当$n \to \infty$时，χ^2分布的极限分布是正态分布

χ^2分布具有可加性

定义 1. $\chi_1^2 \sim \chi^2(n_1)$

2. $\chi_2^2 \sim \chi^2(n_2)$

3. 独立

$\Rightarrow \quad \chi_1^2 + \chi_2^2 \sim \chi^2(n_1 + n_2)$

χ^2分布的数学期望和方差为：

$$E(\chi^2) = n$$

$$D(\chi^2) = 2n$$

6.2.3 t分布

定义 1. $X \sim N(0,1)$

2. $Y \sim \chi^2(n)$

3. X与Y独立

$\Rightarrow \quad t = \frac{X}{\sqrt{Y/n}} \sim t(n)$

t分布的数学期望和方差为：

当$n \geq 2$时，t分布的数学期望$E(t) = 0$

当$n \geq 3$时，t分布的方差为$D(t) = \frac{n}{n-2}$

自由度为1的t分布称为柯西分布，随着自由度的增加，t分布的密度函数越来越接近标准正态分布的密度函数

$n \geq 30$时，认为t分布和标准正态分布一致

与t有关的抽样分布一 设$X_1, X_2, ..., X_n$是来自正态总体$N(\mu, \sigma^2)$的一个样本,

$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$,

$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{x})^2$,

$$\Rightarrow \quad \frac{\sqrt{n}(\overline{X} - \mu)}{S} \sim t(n-1)$$

与t有关的抽样分布二 设X和Y是两个相互独立的总体，$X \sim N(\mu_1, \sigma^2)$，$Y \sim N(\mu_2, \sigma^2)$， $X_1, X_2, ..., X_n$是来自正态总体X的一个样本, $Y_1, Y_2, ..., Y_m$是来自正态总体Y的一个样本, 记：

$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

$\overline{Y} = \frac{1}{m}\sum_{i=1}^{n} Y_i$

$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

$S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$

$S_{xy}^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}$

$$\Rightarrow \quad \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_{xy}\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t(n+m-2)$$

6.2.4 F分布

定义 1. $Y \sim \chi^2(m)$

2. $Z \sim \chi^2(n)$

3. Y与Z独立

$\Rightarrow \quad F = \frac{Y/m}{Z/n} \sim F(m, n)$

F分布的性质：F分布的p分位数$F_p(v_1, v_2)$满足：⇓

$$F_p(v_1, v_2) = \frac{1}{F_{1-p}(v_2, v_1)}$$

与F分布有关的抽样分布 设：

1. $x_1, x_2, ...x_m$是来自$N(\mu_1.\sigma_1^2)$的样本
2. $y_1, y_2, ...y_n$是来自$N(\mu_2.\sigma_2^2)$的样本
3. 两样本相互独立

记号 1. $\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$

2. $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

3. $s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2$

4. $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$

$$\Rightarrow F = \frac{s_x^2/\sigma_1^2}{s_y^2/\sigma_2^2} \sim F(m-1, n-1)$$

F分布与t分布的关系 如果随机变量X服从$t(n)$分布，则X^2服从$F(1, n)$的F分布

6.3 样本均值的分布与中心极限定理

$\overline{X}$的抽样分布 (sampling distribution)仍为正态分布，且

1. $E(\overline{X}) = \mu$
2. $D(\overline{X}) = \frac{\sigma^2}{n}$

   $\Rightarrow \quad \overline{X} \sim N(\mu, \frac{\sigma^2}{n})$

意义 样本均值估计总体，满足无偏性和有效性

中心极限定理 (central limit theorem)

定义 1. 从均值为μ，方差为σ^2(有限)的**任意一个总体**中抽取样本量为n的样本

2. 当n充分大时

   $\Rightarrow$ 样本均值$\overline{X}$近似服从$N(\mu, \frac{\sigma^2}{n})$

意义 不管总体是什么分布，样本均值$\overline{X}$的抽样分布总是近似正态分布，即：

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

在统计学中，由于正态分布有着十分重要的地位，因此把证明其极限分布为正态分布的定理统称为中心极限定理

Chapter 7

特征函数

设$p(x)$是随机变量x的密度函数，则$p(x)$的傅里叶变换为

$$\varphi(t)=\int_{-\infty}^{\infty}e^{itx}p(x)dx$$

7.1 特征函数的定义

设X是一个随机变量，称

$$\varphi(t)=E\left(\mathrm{e}^{\mathrm{i}tx}\right),\quad -\infty<t<\infty$$

为X的特征函数

7.2 利用特征函数求期望方差

$$E(X)=\frac{\varphi'(0)}{\mathrm{i}},\quad Var(X)=-\varphi''(0)+\left(\varphi'(0)\right)^2$$

7.3 特征函数的意义

把寻求**独立随机变量和的卷积运算**(积分运算)转化为乘法运算
把求**分布的各阶原点矩**(积分运算)转换为微分运算
把寻求**随机变量序列的极限分布**转换成一般的函数极限

参考网站：https://www.matongxue.com/madocs/742.html
特征函数是随机变量的分布的不同表现形式
也可以理解为坐标变换，类似于直角坐标和极坐标互换
目的是为了简化运算
思想是：由指数函数e^x的展开式可知，如果随机变量的各阶矩都相等，则其分布相等

$$e^{itX} = 1 + \frac{itX}{1} - \frac{t^2X^2}{2!} + \cdots + \frac{(it)^nX^n}{n!}$$

Chapter 8

参数估计

8.1 参数估计的基本原理

8.1.1 估计量与估计值

参数估计 (parameter estimation) 就是用样本统计量去估计总体的参数

总体参数若用符号θ表示，用于估计总体参数的统计量就用$\hat{\theta}$表示

估计量 (estimator) 用来估计总体参数的统计量，用$\hat{\theta}$表示

比如样本均值、样本比例、样本方差

估计值 (estimated value) 是根据一个样本计算出来的估计量的数值

8.1.2 点估计

点估计 (point estimate) 就是用样本统计量$\hat{\theta}$的某个取值直接作为总体参数θ的估计值

用样本均值$\bar{x}$作为总体均值μ的估计值

用样本比例p作为总体比π的估计值

用样本方差s^2作为总体方差σ^2的估计值

一个点估计的可靠性是由它的抽样标准误差来衡量的

8.1.3 区间估计

区间估计 (interval estimate) 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

区间估计的由来 $\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$

3σ法则

1. 约有68%的数据在平均数±1个标准差的范围内
2. 约有95%的数据在平均数±2个标准差的范围内
3. 约有99%的数据在平均数±3个标准差的范围内

another 3σ

1. 样本均值$\bar{x}$落在总体均值μ的两侧各为1个抽样标准差范围内的概率为0.6827
2. 样本均值$\bar{x}$落在总体均值μ的两侧各为2个抽样标准差范围内的概率为0.9545
3. 样本均值$\bar{x}$落在总体均值μ的两侧各为3个抽样标准差范围内的概率为0.9973

思维转换
如果某个样本的平均值$\bar{x}$落在μ的两个标准差之内
反过来说
μ也就被包括在以$\bar{x}$为中心左右两个标准差的范围内
此处应该补充一个图片，教材128页

置信区间 (confidence interval) 由样本统计量所构造的总体参数的估计区间

置信水平 (confidence level) 置信区间中包含总体参数真值的次数所占的**比例**，也称为置信度或置信系数(confidence confidence)
得到的区间被称为：...的置信度为$1-\alpha$的置信区间
比如对总体比例p的置信区间：总体比例p的置信度为$1-\alpha$的置信区间
此处应该补充一个表格，教材129页

Attention
1. 总体参数的真值是固定的、未知的，而用样本构造的区间是不固定的，置信区间是一个随机区间。就像一个捕鱼的网，不是所有撒网的地点都能捕到鱼
2. 实际问题中，往往只抽取一个样本。主观的希望这个样本能够包含总体参数的真值
3. 不能说置信区间以置信度$1-\alpha$的概率包含总体参数真值
4. 不能说总体参数真值以置信度$1-\alpha$的概率落在置信区间内
5. 真正的意义是：做了无穷次重复抽样，所得到的区间中大约有$1-\alpha\%$包含总体参数。例如做了100次抽样，大概有95次找到的区间包含真值（置信度为0.95）
6. 这个概率是针对随机区间而言的，而不是一个特定的区间

此处应该补充一个图，教材130页

8.1.4 极大似然估计

二项分布的极大似然估计

操作步骤
1. 写出似然函数
$$L(p)=p^{\sum_{i=1}^{n}x_i}(1-p)^{n-\sum_{i=1}^{n}x_i}$$
2. 两端取对数
$$lnL(p)=\sum_{i=1}^{n}x_i ln(p)+\left(n-\sum_{i=1}^{n}x_i\right)(1-p)$$
3. 关于p求偏导，并令其等于零
$$\frac{\partial lnL(p)}{\partial p}=\frac{\sum_{i=1}^{n}x_i}{p}-\frac{n-\sum_{i=1}^{n}x_i}{1-p}=0$$
4. 求出极值点，即p的极大似然估计
$$\hat{p}=\hat{p}(x_1,x_2,...,x_n)=\sum_{i=1}^{n}x_i/n=\overline{x}$$

8.1.5 最小方差无偏估计

8.1.6 评价估计量的标准

问： 什么样的估计量才是一个好的估计量？ ⇓
此处针对每一个特性补充一个图

无偏性 (unbiasedness)是指估计量抽样分布的数学期望等于被估计的总体参数

定义 1. 总体参数θ

2. 估计量$\hat{\theta}$

3. 若：$E(\hat{\theta}) = \theta$

⇒ $\hat{\theta}$为θ的无偏估计量

有效性 (efficiency) 是指同一参数的两个无偏估计量，有更小标准差的估计量更有效。

定义 1. 总体的两个无偏参数$\hat{\theta_1}$ $\hat{\theta_2}$

2. 方差用$D(\hat{\theta_1})$ $D(\hat{\theta_2})$表示

3. 若：$D(\hat{\theta_1}) < D(\hat{\theta_2})$

⇒ $\hat{\theta_1}$是比$\hat{\theta_2}$的更有效的估计量

验证有效性之前必须先验证无偏性

一致性 (consistency)是指随着样本量的增大，估计量的值越来越接近被估计总体的参数

由于样本均值的方差$\sigma_{\overline{x}} = \sigma/\sqrt{n}$，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

8.2 一个总体参数的区间估计

8.2.1 总体均值的区间估计

正态总体、方差已知/非正态总体、大样本 ：

1. 统计量 $z = \frac{\overline{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$
2. 置信水平 $1-\alpha$
3. 点估计值 $\overline{x}$
4. 标准误差 $= z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

$1-\alpha$的置信区间：

$$\overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

正态总体、方差未知、小样本 ：

1. 统计量 $t = \frac{\overline{x}-\mu}{s/\sqrt{n}} \sim t(n-1)$
2. 置信水平 $1-\alpha$
3. 点估计值 $\overline{x}$
4. 标准误差 $= t_{\alpha/2}\frac{s}{\sqrt{n}}$

$1-\alpha$的置信区间：

$$\overline{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$$

8.2.2 总体比例的区间估计

1. 统计量 $z=\frac{p-\pi}{\sqrt{\pi(1-\pi)/n}}\sim N(0,1)$
2. 置信水平 $1-\alpha$
3. 点估计值p
4. 标准误差 $=z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$

$1-\alpha$的置信区间：

$$p\pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

8.2.3 总体方差的区间估计

1. 统计量 $\chi^2=\frac{(n-1)s^2}{\sigma^2}\sim\chi^2(n-1)$
2. 置信水平 $1-\alpha$
3. 点估计值s^2
4. 标准误差　不要求

$1-\alpha$的置信区间：

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}}\leq\sigma^2\leq\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

此处应该补充一个图片，教材138页

8.3 两个总体参数的区间估计

8.3.1 两个正态总体均值之差的区间估计

两个总体均值之差的估计：独立样本

大样本

1. 两个总体独立
2. 两个总体服从正态分布
3. 统计量 $z=\frac{(\overline{x_1}-\overline{x_2})-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}\sim N(0,1)$
4. 置信水平 $1-\alpha$
5. 标准误差 $=z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$

两个总体方差已知

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

两个总体方差未知

$$(\overline{x_1}-\overline{x_2})\pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

小样本 两个总体方差已知

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

假定 1. 两个总体独立

2. 两个总体服从正态分布

3. 两个总体方差未知但相等

4. 统计量 $t=\frac{(\overline{x_1}-\overline{x_2})-(\mu_1-\mu_2)}{s_p\sqrt{1/n_1+1/n_2}}\sim t(n_1+n_2-2)$

⇒ 无论样本量大小，两个样本均值之差都服从正态分布

两个总体方差未知

$s_p^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$

$$(\overline{x_1}-\overline{x_2})\pm t_{\alpha/2}\,(n_1+n_2-2)\,s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

两个总体均值之差的估计：匹配样本

匹配样本 (matched sample) 一个样本中的数据与另一个样本中的数据相对应。

先指定12个工人用第一种方法组装产品，然后再让这12个工人用第二种方法组装产品，这样得到的两种组装产品的数据就是匹配数据，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

大样本 置信区间：

$$\overline{d}\pm z_{\alpha/2}\frac{\sigma_d}{\sqrt{n}}$$

d为两个匹配样本对应数据的差值

$\overline{d}$表示各差值的平均值

σ_d表示各差值的标准差，总体的σ_d未知时可以用样本标准差s_d来代替

小样本 置信区间：

$$\overline{d}\pm t_{\alpha/2}\frac{s_d}{\sqrt{n}}$$

8.3.2 两个总体比例之差的区间估计

两个总体比例之差的抽样分布服从正态分布

统计量$Z=\frac{(p_1-p_2)-(\pi_1-\pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}}\sim N(0,1)$

两个总体比例π_1和π_2未知时，可以用样本比例p_1和p_2来代替

置信区间：

$$(p_1-p_2)\pm z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}$$

8.3.3 两个总体方差比的区间估计

两个方差比的抽样分布 $F=\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}\sim F(n_1-1,n_2-1)$

置信区间：

$$\frac{s_1^2/s_2^2}{F_{\alpha/2}}\le\frac{\sigma_1^2}{\sigma_2^2}\le\frac{s_1^2/s_2^2}{F_{1-\alpha/2}}$$

利用这个性质来求分位数 $F_{\alpha/2}(n_1,n_2)=\frac{1}{F_{1-\alpha/2}(n_2,n_1)}$

8.4 样本量的确定

在一定的样本量下，要提高估计的可靠程度(置信水平)，就应该扩大置信区间，但是过宽的置信区间在实际估计中往往是没有意义的

想要缩小置信区间，又不降低置信度，就需要增加样本量

但是样本量的增加就意味着调查成本的增加

因此要确定一个合理的样本量，使其在满足置信水平下使得调查成本最小

8.4.1 估计总体均值时样本量的确定

在重复抽样或无限总体下，估计的误差是$z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

$z_{\alpha/2}$的值和样本量的值n共同确定了估计误差的大小

置信水平$1-\alpha$确定之后，$z_{\alpha/2}$就确定了

对于给定的$z_{\alpha/2}$值和总体标准差σ，可以确定任一希望的估计误差所需要的样本量

令E代表所希望达到的估计误差，即：$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$，可以得出确定样本量的公式：

$$n = \frac{(z_{\alpha/2})^2\sigma^2}{E^2}$$

E是使用者在给定的置信水平下可以接受的估计误差

样本量的圆整法则： 通常将样本量取值成较大的整数，小数进一

8.4.2 估计总体比例时样本量的确定

在重复抽样或无限总体下，估计的误差是$z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}$

$E = z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}$

确定样本量的公式：

$$n = \frac{(z_{\alpha/2})^2\pi(1-\pi)}{E^2}$$

大多数情况下，取E的值小于0.1

当π的值不知道时，通常取使$\pi(1-\pi)$最大的值0.5

8.5 几个关系

8.5.1 样本量与总体方差的关系

在一定的置信水平下，总体的方差越大，所需要的样本量就越大

考虑一个极端的例子：

总体的方差为0和总体的方差为100

对于总体方差为0的总体，在一定的置信水平下，仅仅需要1个样本就能100%的反应总体的参数

对于总体方差为100的总体，在一定的置信水平下，只有1个样本并不一定能说100%的反应总体的参数

8.5.2 置信水平与置信区间的关系

置信水平可以理解成可以信任的程度

这个值越大，就越值得信任。即我们根据抽样的样本得到的置信区间包含总体参数真值的概率就越大。

举个例子：如果总体的参数是3到7之间的数，那么区间[4,6]和区间[1,100]比起来，区间[1,100]包含总体真值的概率就为1，就比[4,6]值得信任。

从这个例子可以看出：**置信水平越高，置信区间越宽**

8.5.3 估计误差和样本量的关系

描述估计量精度的±值就是估计误差

估计误差也由两部分组成：正态分布的右侧面积为$\alpha/2$的z值和$\frac{\sigma}{\sqrt{n}}$

样本量是包含在估计误差里面的，从公式中可以看出：**样本量的平方和估计误差的大小成反比**

估计误差越大，样本量越小

样本量越大，估计误差越小

8.5.4 置信水平与样本量的关系

样本量是从总体中随机抽取的，抽取的样本量越多，就越能准确的反映总体参数。

考虑一个极端的例子：抽取一个样本量和总体中所有的样本，抽取所有的样本的值一定能反映总体的参数

从这一点可以看出：**样本量越大，就越值得信任，置信水平就越高** 样本量越大，置信水平越高

注意：并不能由样本量越大，置信水平越高，得出置信区间越宽。这几个值之间的关系是：其余量不变，只比较两个值之间的关系

8.6 总结：表格形式

Chapter 9

假设检验

9.1 假设检验的基本问题

9.1.1 假设检验中的P值的意义

1. 在原假设下，检验统计量取其实现值(沿着备选假设方向)更加极端的概率
   如果某检验统计量T的样本实现值为t
   如果T越大就越有利于备选假设，则p值等于原假设下统计量T取其实现值及更极端的概率$P_{H_0}(T \geq t)$
   如果T越小就越有利于备选假设，则p值等于原假设下统计量T取其实现值及更极端的概率$P_{H_0}(T \leq t)$
   如果绝对值$|T|$越大就越有利于备选假设，则p值等于$P_{H_0}(|T| \geq |t|)$

2. 利用样本观测值能够作出的拒绝原假设的最小显著性水平：P值越小，说明这种情况发生的概率很小，而这种情况出现了，根据小概率原理，就有理由拒绝原假设P值

   越小越拒绝
   P值 $< \alpha$，拒绝H_0

9.1.2 假设检验中备择假设的方向

1. 把希望证明的命题放在备择假设中；把原有的、传统的观点或结论放在原假设上。
2. 备择假设的不等式应该按照实际数据代表的方向来确定：

$\overline{x} > \mu_0 \Rightarrow H_1 : \mu > \mu_0$
$\overline{x} < \mu_0 \Rightarrow H_1 : \mu < \mu_0$
通常是被认为可能比零假设更加符合数据所代表的现实
接受备择假设一定意味着原假设错误
没有拒绝原假设不能表明备择假设一定是错的

9.1.3 匹配样本的选择

如果样本可能存在相依关系，选择匹配样本可以提高效率
匹配样本实质上起到了控制观测变量影响因素的作用，可以得到更为精确的推断结果

9.1.4 假设检验也称为显著性检验

计算统计量的过程类似与分数转化过程，如同把一般得分转化为标准得分

9.1.5 假设检验的原理

1. 概率意义上的反证法
   假设H_0 为真，如果发生了与 H_0不一致的、概率小于显著性水平 α的事件，则拒绝H_0
2. 小概率原理：发生概率很小的随机事件在一次试验中几乎不可能发生

第一类错误 α错误：弃真错误，原假设H_0为真却拒绝了

第二类错误 β错误：取伪错误，原假设H_0为伪却没有拒绝

9.1.6 同时使α和β变小的方法

增大样本量

9.1.7 为什么首先控制α错误

原假设是什么通常是明确的 $H_0 : \mu = \mu_0$
备择假设是什么通常是模糊的 $H_1 : \mu < \mu_0 \quad H_1 : \mu > \mu_0 \quad H_1 : \mu \neq \mu_0$

9.1.8 α的缺点

不同的 α，其得出的结果不同，因此需要确定一个最小α，即p

9.1.9 P值大小的影响因素

1. 样本数据与原假设之间的差异
2. 样本量
3. 被假设参数的总体分布

9.1.10 正态性的检验

1. 用Shapiro正态性检验（best）
2. KolmogorovSmirnov检验
3. 正态QQ图（最直观）

9.1.11 总体比例的检验中大样本的确定

当区间

$$p_0 \pm 3\sqrt{\frac{p_0(1-p_0)}{n}}$$

完全包含在$[0,1]$内部时，就近似的认为样本量足够大，能够用正态近似
这里可以理解成另一种3σ原则, 即点估计的值加减3倍标准差仍然在$[0,1]$内

9.1.12 非参数检验

和数据本身的总体分布无关的检验

9.1.13 符号检验

对位置参数中位数的检验

9.1.14 随机性的游程检验

检验取一个或者两个值的变量的这两个值的取值是否是随机的也可以用于某个连续随机变量的取值小于某个确定值及大于该值的个数

9.1.15 接受零假设的不妥

对于同一个检验问题，可能有多种检验方法，但是只要有一个拒绝，就可以拒绝, 那些不能拒绝的检验方法是能力不足, 那些不能拒绝的检验方法是势不足或者效率低

以关于均值的t检验为例，只要零假设的均值和样本均值的确不一样，那么根据检验统计量的公式可以看出，如果样本量不断增大，就必然会拒绝零假设；如果样本量充分小，就必定不能拒绝零假设，这也从另一方面说明了不能说“接受零假设”

9.2 一个总体参数的检验

假设检验 a

1. 一个总体参数——总体均值的检验
2. 正态总体
3. 大样本

已知总体方差

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

未知总体方差，大样本下用样本方差代替

$$z = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

假设检验 a

1. 一个总体参数——总体均值的检验
2. 正态总体
3. 小样本

已知总体方差

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \quad \sim N(0,1)$$

未知总体方差

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \quad \sim t(n-1)$$

假设检验 a

1. 一个总体参数——总体比例的检验
2. 二项分布总体 $np > 5 \quad nq > 5$
3. 大样本

$$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

假设检验 a

1. 一个总体参数——总体方差的检验
2. 正态总体
3. 方差为σ^2

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \quad \sim \chi^2(n-1)$$

假设检验 a

1. 两个总体参数——总体均值差的检验
2. 两个正态总体，或者未知总体分布大样本
3. σ_1^2 σ_2^2已知 $\overline{x_1}-\overline{x_2}$的抽样分布服从正态分布标准差为$\sigma_{\overline{x_1}-\overline{x_2}}=\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$

$$z=\frac{(\overline{x_1}-\overline{x_2})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

9.3 两个总体参数的检验

假设检验 b

1. 两个总体参数——总体均值差的检验
2. 两个正态总体，或者未知总体分布大样本
3. σ_1^2 σ_2^2未知，且n较小
4. $\sigma_1^2=\sigma_2^2$ $\overline{x_1}-\overline{x_2}$的抽样分布服从$t$分布 $\sigma_{\overline{x_1}-\overline{x_2}}$的估计为 $\hat{\sigma}_{\overline{x_1}-\overline{x_2}}=s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$
$s_p^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$

$$t=\frac{(\overline{x_1}-\overline{x_2})-(\mu_1-\mu_2)}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\quad\sim t(n_1+n_2-2)$$

假设检验 b

1. 两个总体参数——总体均值差的检验
2. 两个正态总体，或者未知总体分布大样本
3. σ_1^2 σ_2^2未知，且n较小
4. $\sigma_1^2\neq\sigma_2^2$ $\sigma_{\overline{x_1}-\overline{x_2}}$ 的估计为 $\hat{\sigma}_{\overline{x_1}-\overline{x_2}}=\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$ $f=\frac{(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2})^2}{\frac{(\frac{s_1^2}{n_1})^2}{n_1-1}+\frac{(\frac{s_2^2}{n_2})^2}{n_2-1}}$

$$t=\frac{(\overline{x_1}-\overline{x_2})-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}\quad\sim t(f)$$

假设检验 b

1. 两个总体参数——总体比例之差的检验
2. 两个二项分布总体
3. 大样本
4. 总体比例为π
5. 样本比例为 $p=\frac{x}{n}$ $p=\frac{x_1+x_2}{n_1+n_2}=\frac{p_1n_1+p_2n_2}{n_1+n_2}$

$$z=\frac{p_1-p_2}{\sqrt{p(1-p)(\frac{1}{n_1}+\frac{1}{n_2})}}\quad\sim N(0,1)$$

$$z=\frac{(p_1-p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}}\quad\sim N(0,1)$$

假设检验 b

1. 两个总体参数——总体比例的检验
2. 两个二项分布总体
3. 大样本
4. 总体比例为π
5. 样本比例为$p=\frac{x}{n}$

6. 检验比例之差不为零

$$z=\frac{(p_1-p_2)-(\pi_1-\pi_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}}=\frac{(p_1-p_2)-d_0}{\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}}\quad\sim N(0,1)$$

假设检验 b

1. 两个总体参数——总体方差比的检验
2. 两个正态总体 $F=\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}\quad\sim F(n_1-1,n_2-1)$ 临界点

$$\{F_{1-\alpha/2}(n_1-1,n_2-1)\quad,\quad F_{\alpha/2}(n_1-1,n_2-1)\}$$

9.4 课后习题

9.4.1 假设检验和参数估计有什么相同点和不同点

相同点：都是利用样本统计量对总体参数进行某种判断，都是以抽样分布为理论依据，都是建立在概率上的推断，推断结果都有一定的可信度或风险

不同点：推断的角度不同

参数估计：讨论的是样本统计量估计总体参数的方法，总体参数μ在估计前是未知的

假设检验：先对μ的值提出一个假设，然后利用样本信息去检验这个假设是否成立

9.4.2 什么是假设检验中的显著性水平？

假设检验中的显著性水平：显著性水平是当原假设正确时却被拒绝的概率或风险，即假设检验中犯弃真错误的概率，通常用α表示

注意：参数估计中有置信水平,假设检验中有显著性水平

9.4.3 统计显著是什么意思？

统计显著：统计显著是指在原假设为真的条件下，用于检验的样本统计量的值落在了拒绝域内，做出了拒绝原假设的决定。但是这并不意味着犯第一类错误，因为用于检验的样本统计量只有在假定原假设为真的情况下才能计算出来

9.4.4 什么是假设检验中的两类错误？

第一类错误：原假设H_0为真却拒绝了—— α错误——弃真错误

第二类错误：原假设H_0为伪却没有拒绝—— β错误——取伪错误

9.4.5 假设检验中的两类错误有什么数量关系?

如果减小α错误，就会增大犯β错误的机会

如果减小β错误，就会增大犯α错误的机会

如果想使得α和β错误同时变小，就只有增大样本量

9.4.6 解释假设检验中的P值

1. 在原假设成立下，根据样本计算出的统计量取极端值的概率
2. 利用样本观测值能够作出的拒绝原假设的最小显著性水平

P值越小，说明这种情况发生的概率很小，而这种情况出现了，根据小概率原理，就有理由拒绝原假设P值越小越拒绝

P值 $<\alpha$，拒绝H_0

9.4.7 显著性水平与P值的区别

显著性水平：显著性水平是当原假设正确时却被拒绝的概率或风险，即假设检验中犯弃真错误的概率，通常用α表示

P值：在原假设成立下，根据样本计算出的统计量取极端值的概率；利用样本观测值能够作出的拒绝原假设的最小显著性水平

P值越小，说明这种情况发生的概率很小，而这种情况出现了，根据小概率原理，就有理由拒绝原假设
P值越小越拒绝
P值 $< \alpha$，拒绝H_0

区别：

1. 显著性水平是当原假设正确时却被拒绝的概率或风险；
   P值是在原假设成立下，根据样本计算出的统计量取极端值的概率

2. 显著性水平是人们根据检验的要求确定的，表示的是犯第一类错误的概率；
   P值是通过计算得到的，是拒绝原假设的最小显著性水平，比一般的显著性水平更加精确

9.4.8 P值大小取决于三个因素：

1. 样本数据与原假设之间的差异
2. 样本量
3. 被假设参数的总体分布

9.4.9 假设检验依据的原理是什么

1. 概率意义上的反证法：假设H_0 为真，如果发生了与 H_0不一致的、概率小于显著性水平 α的事件，则拒绝H_0
2. 小概率原理：发生概率很小的随机事件在一次试验中几乎不可能发生

9.4.10 单侧检验中原假设和备择假设的方向应该如何确定？

单侧检验的两种情况：
考察的数值越大越好——左单侧检验/下限检验
考察的数值越小越好——右单侧检验/上限检验

Chapter 10

分类数据分析

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
χ^2统计量

$$\chi^2 = \Sigma \frac{(f_0 - f_e)^2}{f_e}$$

变量含义：
1. f_0：观察值频数
2. f_e：期望值频数

特征：
1. $\chi^2 \geq 0$
2. χ^2统计量的分布与自由度有关
3. χ^2统计描述了观察值与期望值的接近程度；越接近，χ^2统计量值越小

10.1 拟合优度检验是什么

用χ^2统计量进行统计显著性检验
1. 根据总体分布状况;
2. 计算出分类变量中各类别的期望频数;
3. 与分布的观察频数进行对比;
4. 判断期望频数与观察频数是否有差异;
5. 从而达到对分类变量进行分析的目的.

10.1.1 χ^2拟合优度检验的步骤：

例题： 1912年4月15日，泰坦尼克号与冰山相撞，当时船上共有2208人，男性1738人，女性470人
海难发生后，幸存者共718人，男性374人，女性344人。
$\alpha = 0.1$的显著性检验水平检验存活状况与性别是否相关
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
表中第一行数据为男性数据，期望频数$f_e = \frac{1738}{2208} * 718 = 153$
表中第二行数据为女性数据，期望频数$f_e = \frac{470}{2208} * 718 = 153$

1. 计算$f_0 - f_e$

2. 计算$(f_0 - f_e)^2$

3. 计算$\frac{(f_0-f_e)^2}{f_e}$

4. 计算$\chi^2 = \Sigma\frac{(f_0-f_e)^2}{f_e} = 303$

   自由度的计算公式为$df = R - 1$

R为分类变量类型的个数，本例中为两个性别
结果得出，χ^2远大于$\chi^2_{0.1}$，则拒绝原假设，认为观察频数与期望频数不一致

10.2 列联分析

拟合优度检验是对一个分类变量的检验，对两个分类变量的分析，称为：独立性检验

10.2.1 列联表

将两个以上的变量进行交叉分析的频数分布表
行(Row)，记为R
列(Column)，记为C
这样，列联表就叫做RxC列联表

10.2.2 独立性检验

分析列联表中行变量和列变量是否相互独立
独立性检验：例题
一种原料来自三个不同的地区，原料质量被分为三个不同的等级。从这批原料中随机抽取500件进行检验，结果如下所示，要求检验各个地区和原料等级之间是否存在依赖关系($\alpha = 0.05$)

地区	一级	二级	三级	合计
甲	52	64	24	140
乙	60	59	52	171
丙	50	65	74	189
合计	162	188	150	500

H_0：地区和原料之间是独立的（不存在依赖关系）
H_1：地区和原料之间不独立（存在依赖关系）
分析的关键是获得期望值
第一行，甲地区的合计为140，用140/500作为甲地区原料比例的估计值
第一列，一级原料的合计为162，用162/500作为一级原料比例的估计值
如果独立，则满足：
令A：样本单位来自甲地区的事件
B：样本单位属于一级原料的事件
$P(c_{11}) = P(AB)$
$= P(A)P(B)$
$= (\frac{140}{500})(\frac{162}{500})$
$= 0.09072$ 0.09072是第一个单位中的期望比例，相应的频数期望值为$0.09072 * 500 = 45.36$
采用下式来计算任何一个单元中频数的期望值

$$f_e = \frac{RT}{n} * \frac{CT}{n} * n = \frac{RT * CT}{n}$$

f_e为给定单元中的频数值
RT为给定单元中的行合计

CT为给定单元中的列合计

n为样本量

如下表：

χ^2的自由度$=(R-1)(C-1)=2*2=4$

令$\alpha=0.05$，查表知：$\chi^2_{0.05}(4)=9.488$

因为$\chi^2>\chi^2_{0.05}(4)$，故拒绝H_0，认为地区和原料之间存在依赖关系

10.2.3 分类数据中的相关称为：

品质相关

10.2.4 怎样测量分类数据的相关程度

10.3 列联表中的相关测量

1. φ相关系数

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

2. c系数（列联相关系数）

$$c=\sqrt{\frac{\chi^2}{\chi^2+n}}$$

3. V相关系数

$$V=\sqrt{\frac{\chi^2}{n*min[(R-1),(C-1)]}}$$

10.3.1 φ相关系数

描述2×2列联表数据相关程度最常用的一种相关系数

计算公式：

$$\varphi=\sqrt{\chi^2/n}$$

对于2×2列联表，计算出的φ系数可以控制在$0\sim1$之间

另一个简化公式：

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

当$|\varphi|=1$时，必有某个方向对角线上的值全为0

10.3.2 c相关系数（列联系数）

主要用于列联表大于2×2的情况

计算公式：

$$c=\sqrt{\frac{\chi^2}{\chi^2+n}}$$

特点：

1. 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

2. 不便于比较不同行列的两个列联系数

10.3.3 V相关系数

鉴于φ系数无上限，c系数小于1的情况，给出了V相关系数

计算公式：

$$V=\sqrt{\frac{\chi^2}{n*min[(R-1),(C-1)]}}$$

取值为$0\sim1$

10.4 在对不同的列联表变量之间的相关程度进行比较时应该注意的问题

1. 不同列联表中行与行、列与列的个数要相同
2. 采用同一种系数

10.4.1 χ^2分布的期望值准则

用χ^2分布来进行独立性检验，要求样本量必须足够大，特别是每个单元中的期望频数（理论频数）不能过小，否则会得出错误的结论

准则一：如果只有两个单元，则每个单元的期望频数必须是5或以上

准则二：倘若有两个以上的单元，20%的单元的期望频数$f_e<5$，则不能使用χ^2检验

10.5 课后习题

10.5.1 简述列联表的构造和列联表的分布

列联表是将两个以上的变量进行交叉分类的频数分布表

列联表的分布观察值的分布——条件分布（每个具体的观察值就是条件频数）期望值的分布

10.5.2 说明计算χ^2统计量的步骤

计算f_0-f_e

计算$(f_0-f_e)^2$

计算$\frac{(f_0-f_e)^2}{f_e}$

计算$\chi^2=\Sigma\frac{(f_0-f_e)^2}{f_e}$

10.5.3 简述φ系数、c系数、V系数的各自特点

1. φ相关系数

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

适用于$2x2$列联表,取值范围为$[0,1]$

两个分类变量完全相关，则$|\varphi|=1$，且$2x2$列联表中某一方向对角线上的值全为0

2. c系数（列联相关系数）

$$c=\sqrt{\frac{\chi^2}{\chi^2+n}}$$

适用于大于$2x2$列联表, 两个变量独立，$c=0$

特点：可能的最大值依赖于列联表中的行数和列数，且随着R和C的增大而增大

c系数的值小于φ系数

3. V相关系数

$$V=\sqrt{\frac{\chi^2}{n*min[(R-1),(C-1)]}}$$

两个变量独立，$V=0$;
两个变量完全相关，$V=1$

例题 从总体中随机抽取$n=200$的样本，调查后按不同属性归类，得到如下结果：

$$n_1=28, n_2=56, n_3=48, n_4=36, n_5=32$$

依据经验数据，各类别在总体中的比例分别为：

$$\pi_1=0.1, \pi_2=0.2, \pi_3=0.3, \pi_4=0.2, \pi_5=0.2$$

以$\alpha=0.1$的显著性水平进行检验，说明现在的情况与经验数据相比是否发生了变化（用P值）
1. 计算f_0-f_e
2. 计算$(f_0-f_e)^2$
3. 计算$\frac{(f_0-f_e)^2}{f_e}$
4. 计算$\chi^2=\Sigma\frac{(f_0-f_e)^2}{f_e}$ 计算的时候期望频数应该尽量取小数点后几位，这样算出来更加精确

按照标准的四步法计算出来的$\chi^2=14$
P值为：

$$P\{\chi^2(5-1)>14\}=0.007295<0.1=\alpha$$

P值是在原假设成立下，根据样本计算出的统计量取极端值的概率

在原假设下，检验统计量取其实现值(沿着备选假设方向)更加极端的概率
如果某检验统计量T的样本实现值为t
如果T越大就越有利于备选假设，则p值等于原假设下统计量T取其实现值及更极端的概率$P_{H_0}(T\geq t)$
如果T越小就越有利于备选假设，则p值等于原假设下统计量T取其实现值及更极端的概率$P_{H_0}(T\leq t)$
如果绝对值$|T|$越大就越有利于备选假设，则p值等于$P_{H_0}(|T|\geq|t|)$

Chapter 11

方差分析

11.1 方差分析引论

从形式上看，方差分析是什么：从形式上看，方差分析是比较多个总体的均值是否相等

从本质上看，方差分析是什么：研究变量之间的关系。研究一个（或多个）分类型自变量与一个数值型因变量之间的关系

方差分析 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

在方差分析中，所要检验的对象称为 因素或因子，比如说行业

在方差分析中，因素的不同表现称为 水平或处理，比如行业的不同表现有：零售业、旅游业、航空公司、家电制造业等等

在方差分析中，在每个因子水平下得到的样本数据称为 观测值

在只有一个因素的方差分析（单变量方差分析）中，涉及两个变量 分类型自变量，数值型自变量

11.1.1 为什么检验均值是否相等的分析叫做方差分析？和分析方差有什么关系吗？

虽然人们感兴趣的是均值，但是判断均值之间是否有差异时，需要借助于方差。这个名字也表明，它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。进行方差分析时，需要考虑数据误差的来源（组内误差、组间误差）

组内误差 Factor A 来自水平内部的数据误差，比如零售业中的抽样数据的抽样误差。组内误差只含有随机误差，组内误差主要是由抽样误差导致的

组间误差 Error 不同水平之间的数据误差，比如零售业和旅游业之间的数据误差。组间误差是随机误差和系统误差的总和

组内平方和 SSE： sum of squares for error 反映组内误差大小（随机误差/抽样误差）反映了每个样本内各观测值的离散状况，是随机误差大小的度量。反映了除自变量对因变量的影响之外其他因素对因变量的总影响，也称为残差变量，引起的误差称为残差效应

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{t_i} (x_{ij} - \overline{x_i})^2$$

简化计算式：$S_e = S_T - S_A$

组间平方和 SSA： sum of squares for factor A 反映组间误差大小的平方和，反映了样本均值之间的差异程度，是随机误差和系统误差大小的度量，反映了自变量（行业）对因变量（被投诉次数）的影响，被称为自变量效应或因子效应

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

简化计算式：

$$S_A = \sum_{i=1}^{k} \frac{Y_{i\cdot}^2}{t_i} - n\overline{Y}^2$$

推导：

$$\begin{aligned} SSA &= \textstyle\sum_{i=1}^{k} n_i \left(\overline{x_i} - \overline{\overline{x}}\right)^2 \\ &= \textstyle\sum_{i=1}^{k} n_i \left(\overline{x_i}^2 - 2 * \overline{x_i} * \overline{\overline{x}} + \overline{\overline{x}}^2\right) \\ &= \textstyle\sum_{i=1}^{k} n_i \overline{x_i}^2 - \sum_{i=1}^{k} 2 * \overline{x_i} * \overline{\overline{x}} + \sum_{i=1}^{k} \overline{\overline{x}}^2 \\ &= \textstyle\sum_{i=1}^{k} n_i \overline{x_i}^2 - 2n\overline{\overline{x}}^2 \\ &= \textstyle\sum_{i=1}^{k} \frac{x_i^2}{n_i} - 2n\overline{\overline{x}}^2 \end{aligned}$$

总平方和 SST： sum of squares for total 反映全部数据误差大小的平方和反映了全部观测值的离散状况是对全部数据总误差程度的度量反映了自变量和残差变量的共同影响，等于残差效应 + 自变量效应

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

简化计算式：$S_T = \sum_{i=1}^{k} \sum_{j=1}^{t} Y_{ij}^2 - n\overline{Y}^2$

推导：

$$\begin{aligned} S_T &= \textstyle\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \overline{\overline{x}}\right)^2 \\ &= \textstyle\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij}^2 - 2 * x_{ij} * \overline{\overline{x}} + \overline{\overline{x}}^2\right) \\ &= \textstyle\sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}^2 - \sum_{i=1}^{k} \sum_{j=1}^{n_i} 2 * x_{ij} * \overline{\overline{x}} + \sum_{i=1}^{k} \sum_{j=1}^{n_i} \overline{\overline{x}}^2 \\ &= \textstyle\sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}^2 - 2n\overline{\overline{x}}^2 \end{aligned}$$

误差分析的基本思想 组间误差比组内误差

组间误差（随机误差 + 系统误差）

组内误差（随机误差）

查看系统误差占比多少，占比比较大，就认为检验显著

11.1.2 方差分析中的基本假定

每个总体服从正态分布

每个总体的方差σ^2必须相同

观测值独立

11.1.3 方差分析分类

（根据所分析的分类型自变量的多少）

单因素方差分析(one-way analysis of variance):方差分析中只涉及一个分类型自变量

双因素方差分析(two-way analysis of variance):方差分析中涉及两个分类型自变量

11.2 单因素方差分析

组间均方/组间方差 MSA SSA的均方自由度为$k-1$

$$MSA = \frac{SSA}{k-1}$$

组内均方/内间方差 MSE SSE的均方自由度为$n-k$

$$MSE=\frac{SSE}{n-k}$$

检验统计量F

$$F=\frac{MSA}{MSE}\quad\sim\quad F(k-1,n-k)$$

11.2.1 方差分析的拒绝域为什么是右单侧

H_0成立时，希望S_A越小越好，因此拒绝域也只有单边的

11.2.2 方差分析中关系强度的度量

用组间平方和SSA占总平方和SST的比例大小来反映，记为 R^2

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

表明自变量（行业）对因变量（被投诉次数）的影响效应占总效应的$100*R^2\%$

11.2.3 方差分析中的多重比较

如果拒绝了H_0，则需要找出是哪几个水平之间的因素不同
通过对总体均值之间的配对比较来检验到底哪些均值之间存在差异
费希尔提出了LSD方法（最小显著差异方法）

11.2.4 方差分析中的多重比较 LSD方法

步骤：
提出假设：$H_0:\mu_i=\mu_j\quad H_1:\mu_i\neq\mu_j$
计算检验统计量：$\overline{x}_i-\overline{x}_j$
计算LSD，t统计量的自由度为 $n-k$：

$$LSD=t_{\alpha/2}(n-k)\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

根据显著性水平α作出决策：
如果 $|x_i-x_j|>LSD$，则拒绝 H_0
如果 $|x_i-x_j|<LSD$，则不拒绝 H_0

Chapter 12

一元线性回归

12.1 变量间关系的度量

变量间的关系可分为 函数关系、相关关系

函数关系一一对应的确定关系

相关关系(correlation) 变量之间存在的不确定的数量关系

进行相关分析时，对总体的假定： 两个变量之间是线性关系、两个变量都是随机变量

12.1.1 相关系数

(Pearson's correlation coefficient)(Pearson相关系数)

$$r=\frac{n\Sigma xy-\Sigma x\Sigma y}{\sqrt{n\Sigma x^2-(\Sigma x)^2}*\sqrt{n\Sigma y^2-(\Sigma y)^2}}$$

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，取值范围[-1,1]

1. r具有对称性
2. r的数值大小与x和y的原点及尺度无关
3. r仅仅是x与y之间线性关系的一个度量，它不能用于描述非线性关系

$$r=\sqrt{R^2}=\sqrt{\hat{\beta}_1\frac{l_{xy}}{l_{yy}}}=\sqrt{\frac{l_{xy}^2}{l_{xx}l_{yy}}}$$

总体相关系数ρ是未知的，需要用样本相关系数r作为ρ的近似值。因此，r是一个随机变量

r的抽样分布 随着总体相关系数ρ和样本量的大小而变化

样本数据接近正态总体，n增大，r的抽样分布接近正态分布：

ρ很小或接近0时，趋于正态分布的趋势十分明显

ρ远离0时，除非n非常大，否则r的抽样分布呈现一定的偏态

ρ为较大的正值时，r呈现左偏分布

ρ为较大的负值时，r呈现右偏分布

ρ接近于0，样本量n很大，才能认为r是接近正态分布的随机变量

12.1.2 r的显著性检验

若r服从正态分布假设，则应该用正态性检验，但是对r假定为正态分布有风险，所以通常情况下不采用正态分布检验

ρ是总体相关系数

检验方法：

t检验适用范围：大样本、小样本

1. 提出假设： $H_0 : \rho = 0; \quad H_1 : \rho \neq 0$
2. 计算检验统计量：$t = |r|\sqrt{\frac{n-2}{1-r^2}} \sim t(n-2)$
3. 进行决策：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

注意：即使统计检验表明相关系数在统计上是显著的，也并不意味着两个变量之间就存在重要的相关性。因为在大样本情况下，几乎总是导致相关系数显著。样本大，计算出的t值就很大，就容易落入拒绝域

12.2 一元线性回归

12.2.1 回归分析解决的问题：

从一组数据出发，确定变量之间的数学关系式对这些关系书的可信程度进行各种统计检验，并从影响某一特定变量的诸多变量中找出哪些变量的影响是显著的，哪些是不显著的

利用所求的关系式，根据一个或几个变量的取值来估计或预测另一个特定变量的取值，并给出这种估计或预测的可靠程度

自变量(independent variable) 用来预测或解释因变量的一个或多个变量

因变量(dependent variable) 被预测或被解释的变量

回归模型(regression model) 描述因变量y如何依赖自变量x和误差项ε的方程只涉及一个自变量的一员线性回归模型可表示为：

$$y = \beta_0 + \beta_1 x + \varepsilon$$

$\beta_0 + \beta_1 x$反映了由于x的线性变化而引起的y的线性变化

ε是被称为误差项的随机变量，反映了除x和y之间的线性关系之外的随机因素对y的影响，是不能由x和y之间的线性关系所解释的变异性

回归模型(regression model)的假定 a

对误差项的假定：

1. 误差项ε是一个期望值为0的随机变量，即$E(\varepsilon) = 0$
2. 对于所有的x值，ε的方差σ^2都相同
3. $\varepsilon \sim N(0,1)$，ε独立，独立性意味着一个特定的x值所对于的ε与其他x值所对应的ε不相关

对自变量x和因变量y的假定：

1. 自变量x和因变量y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是非随机的对于任何的一个给定的x值
3. $y \sim N(\beta_0 + \beta_1 x, \sigma^2)$且对于不同的x具有相同的方差

回归方程(regression equation) 描述因变量y的期望值如何依赖于自变量x的方程

一元线性回归方程的形式：

$$E(y) = \beta_0 + \beta_1 x$$

估计的回归方程：用样本统计量$\hat{\beta}_0$和$\hat{\beta}_1$来代替回归方程中的未知参数β_0和β_1

一元线性回归，估计的回归方程形式为：

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

12.2.2 参数的最小二乘估计思想

由高斯提出对于第i个值，估计的回归方程可表示为：

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

最小二乘法：找出一条直线，使其与其余的点的距离最小(垂直距离)

12.2.3 参数的最小二乘估计方法

使

$$\Sigma(y_i - \hat{y}_i)^2 = \Sigma(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

最小

即分别对β_0和β_1求其偏导数，令其等于0

$$\left.\frac{\partial Q}{\partial \beta_0}\right|_{\beta_0=\hat{\beta}_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\left.\frac{\partial Q}{\partial \beta_1}\right|_{\beta_1=\hat{\beta}_1} = -2\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

解上述方程组得：

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

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

当$x = \overline{x}$时，$\hat{y} = \overline{y}$，即回归直线$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$通过点$(\overline{x}, \overline{y})$

12.2.4 一元线性回归算系数的时候的简便算法

令：

$\overline{x} = \frac{1}{n}\Sigma x_i \quad , \quad \overline{y} = \frac{1}{n}\Sigma y_i$

$l_{xy} = \Sigma(x_i - \overline{x})(y_i - \overline{y}) = \Sigma x_i y_i - \frac{1}{n}\Sigma x_i \Sigma y_i = \underline{\Sigma x_i y_i - n\overline{x}\overline{y}}$

$l_{xx} = \Sigma(x_i - \overline{x})^2 = \Sigma x_i^2 - \frac{1}{n}(\Sigma x_i)^2 = \underline{\Sigma x_i^2 - n\overline{x}^2}$

$l_{yy} = \Sigma(y_i - \overline{y})^2 = \Sigma y_i^2 - \frac{1}{n}(\Sigma y_i)^2 = \underline{\Sigma y_i^2 - n\overline{y}^2}$

可得：

$\hat{\beta}_1 = l_{xy}/l_{xx}$

$\hat{\beta}_0 = \overline{y} - \hat{\beta}_0 \overline{x}$

当$x = \overline{x}$时，$\hat{y} = \overline{y}$，

即回归直线$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i x_i$通过点$(\overline{x}, \overline{y})$

12.3 回归直线的拟合优度(goodness of fit)

拟合优度 1. 判定系数 $R^2=\frac{\hat{\beta}_1 l_{xy}}{l_{yy}}=\frac{l_{xy}^2}{l_{xx}l_{yy}}$

2. 标准误差$s_e=\sqrt{MSE}$

回归直线与各观测点的接近程度

12.3.1 回归直线的拟合优度(goodness of fit)的判定系数(coefficient of determination)

y取值的波动称为变差：自变量x的取值不同造成的除了x以外的其余因素

离差$(y-\overline{y})$

n次观测值的总变差可以用这些离差的平方和来表示，称为总平方和SST:

$$SST=\Sigma(y_i-\overline{y})^2$$

离差分解:

$$y-\overline{y}=(y-\hat{y})+(\hat{y}-\overline{y})$$

$$\Sigma(y_i-\overline{y})^2=\Sigma(y_i-\hat{y}_i)^2+\Sigma(\hat{y}_i-\overline{y})^2$$

$\Sigma(y_i-\hat{y}_i)^2$是回归值$\hat{y}_i$与均值$\overline{y}$的离差平方和，反映了y的总变差中由于x和y之间的线性关系引起的变化部分，是可以由回归直线来解释的y_i的变差部分，称为回归平方和，记为SSR。

$\Sigma(\hat{y}_i-\overline{y})^2$是各实际观测点与回归值的残差$(y_i-\hat{y}_i)$的平方和，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y_i的变差部分，称为残差平方和或误差平方和，记为SSE。

SSR: sum of squares for factor regression

SSE: sum of squares for error

总平方和SST=回归平方和SSR + 残差平方和SSE

判定系数(coefficient of determination)，记为R^2

$$R^2=\frac{SSR}{SST}=\frac{\Sigma(\hat{y}-\overline{y})^2}{\Sigma(y_i-\overline{y})^2}=1-\frac{\Sigma(y_i-\hat{y}_i)^2}{\Sigma(y_i-\overline{y})^2}=\frac{\hat{\beta}_1 l_{xy}}{l_{yy}}=\frac{l_{xy}^2}{l_{xx}l_{yy}}$$

$R^2=1$，拟合是完全的，所有观测点都落在直线上

$R^2=0$，x完全无助于解释y的变差取值范围[0,1]

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

判定系数的实际意义： 有$R^2\%$可以由x和y的线性关系来解释

12.3.2 回归直线的拟合优度(goodness of fit)的标准误差(standard error of estimate)

判定系数可以用来度量回归直线的拟合优度

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

估计标准误差就是度量各实际观测点在直线周围的散布状况的一个统计量,用s_e来表示

$$s_e=\sqrt{\frac{\Sigma(y_i-\hat{y}_i)^2}{n-2}}=\sqrt{\frac{SSE}{n-2}}=\sqrt{MSE}$$

估计标准误差是对误差项ε的标准差 σ的估计

回归直线是对n各观测点拟合的所有直线中，估计标准误差最小的一条直线

12.3.3 显著性检验——线性关系检验

F检验 $y=\beta_0+\beta_1 x+\varepsilon$

1. 提出假设H_0： $\beta_1=0$ (两个变量之间线性关系不显著)
2. 计算检验统计量F
$$F=\frac{SSR/1}{SSE/(n-2)}=\frac{MSR}{MSE} \quad \sim \quad F(1,n-2)$$
3. 作出决策：
$F>F_\alpha$，拒绝H_0，表明两个变量之间的线性关系是显著的$F<F_\alpha$，不拒绝H_0

$$S_R=\hat{\beta}^2 l_{xx}=\hat{\beta} l_{xy}$$
$$S_e=l_{yy}-S_R=l_{yy}-\hat{\beta} l_{xy}$$

12.3.4 显著性检验——回归系数检验

t检验 $y=\beta_0+\beta_1 x+\varepsilon$ 检验 β_1

1. 提出假设：H_0： $\beta_1=0$ $\beta_1\neq 0$
2. 计算检验统计量t
$$t=\frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}\sim t(n-2)$$
3. 做出决策：
$|t|>t_{\alpha/2}$，拒绝H_0，表明自变量x对因变量y的影响是显著的
$|t|<t_{\alpha/2}$，不拒绝H_0

符号说明：
$$s_e=\sqrt{\frac{\Sigma(y_i-\hat{y}_i)^2}{n-2}}=\sqrt{\frac{SSE}{n-2}}=\sqrt{MSE}$$
$$s_{\hat{\beta}_1}=\frac{s_e}{\sqrt{\Sigma x_i^2-\frac{1}{n}(\Sigma x_i)^2}}$$
$$t=\frac{\hat{\beta}_1-\beta_1}{s_{\hat{\beta}_1}}\sim t(n-2)$$

在一元线性回归中，线性关系检验和回归系数检验是等价的

在多元线性回归中，线性关系检验是检验回归方程的线性关系是否显著，回归系数检验是检验回归方程的回归系数是否显著

如果线性关系检验显著，那么就说明x和y可以用一个方程来表示，F检验是用来检验总体的回归关系的显著性

如果回归系数检验显著，那么就说明其对应的偏回归系数β_i不为0，t检验是用来检验各回归系数的显著性

参考多元线性回归部分的显著性检验

12.4 利用回归方程进行预测

12.4.1 点估计

利用估计的回归方程，对于x的一个特定值x_0，求出y的一个估计值就是点估计

平均值的点估计：利用估计的回归方程，对于x的一个特定值x_0，求出y的平均值的一个估计值$E(y_0)$ （总体参数）

$$E(y_0)=\beta_0+\beta_1 x_0$$

个别值的点估计：利用估计的回归方程，对于x的一个特定值x_0，求出y的一个个别值的估计值y_0 （具体取值）

$$\hat{y_0} = \beta_0 + \beta_1 x_0$$

在点估计下，对于同一个x_0，平均值的点估计和个别值的点估计的结果是一样的，但在区间估计中则有所不同

12.4.2 区间估计

利用估计的回归方程，对于x的一个特定值x_0，求出y的一个估计值的区间
置信区间估计：对于x的一个给定值x_0，求出y的平均值的一个估计区间
预测区间估计：对于x的一个给定值x_0，求出y的一个个别值的估计区间

12.4.3 区间估计——置信区间估计（y的平均值的置信区间估计）

对于x的一个给定值x_0，求出y的平均值的一个估计区间
设： x_0为自变量x的一个特定值或给定值$E(y_0)$为给定x_0时因变量y的平均值或期望值
当$x = x_0$时，$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$为$E(y_0)$的估计值
一般来说，不能期望估计值$\hat{y_0}$精确的等于$E(y_0)$
$\hat{y_0}$的标准差的估计量为$s_{\hat{y_0}}$：

$$s_{\hat{y_0}} = s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\Sigma(x_i - \overline{x})^2}}$$

置信区间：

$$\hat{y_0} \pm t_{\alpha/2} s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\Sigma(x_i - \overline{x})^2}}$$

12.4.4 区间估计——预测区间估计（y的个别值的置信区间估计）

对于x的一个给定值x_0，求出y的一个个别值的估计区间
y的一个个别值的标准差的估计量s_{ind}：

$$s_{ind} = s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\Sigma(x_i - \overline{x})^2}}$$

置信区间：

$$\hat{y_0} \pm t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\Sigma(x_i - \overline{x})^2}}$$

12.4.5 区间估计——置信区间估计和预测区间估计的比较

置信区间估计：

$$\hat{y_0} \pm t_{\alpha/2} s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\Sigma(x_i - \overline{x})^2}}$$

预测区间估计：

$$\hat{y_0} \pm t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\Sigma(x_i - \overline{x})^2}}$$

从公式和下图可看出，预测区间比置信区间要宽

12.4.6 标准化残差

$$z_e = \frac{e_i}{s_e} = \frac{y_i - \hat{y}_i}{s_e}$$

Chapter 13

一元线性回归重要公式、多元线性回归

13.1 一元线性回归重要公式

13.1.1 相关系数

$$r = \frac{n\Sigma xy - \Sigma x\Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} * \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

1. 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
2. 取值范围[-1,1]
3. r具有对称性
4. r的数值大小与x和y的原点及尺度无关
5. r仅仅是x与y之间线性关系的一个度量，它不能用于描述非线性关系

r是样本相关系数： ρ 是总体相关系数

1. 提出假设： $H_0 : \rho = 0; \quad H_1 : \rho \neq 0$
2. 计算检验统计量：$t = |r|\sqrt{\frac{n-2}{1-r^2}} \sim t(n-2)$
3. 进行决策：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

13.1.2 一元线性回归

$E(y) = \beta_0 + \beta_1 x$
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
方法：最小二乘法

1. 使$\Sigma(y_i - \hat{y}_i)^2 = \Sigma(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$最小
2. 即分别对β_0 和β_1求其偏导数，令其等于0
   $\left.\frac{\partial Q}{\partial \beta_0}\right|_{\beta_0=\hat{\beta}_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$
   $\left.\frac{\partial Q}{\partial \beta_1}\right|_{\beta_1=\hat{\beta}_1} = -2\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$
3. 解上述方程组得：
   $\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$
   $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$

4. 当$x=\overline{x}$ 时，$\hat{y}=\overline{y}$ ，即回归直线$\hat{y_i}=\hat{\beta}_0+\hat{\beta}_1 x_i$ 通过点$(\overline{x},\overline{y})$

另一种简便方法：

令：

$\overline{x}=\frac{1}{n}\Sigma x_i \quad , \quad \overline{y}=\frac{1}{n}\Sigma y_i$

$l_{xy}=\Sigma(x_i-\overline{x})(y_i-\overline{y})=\Sigma x_i y_i-\frac{1}{n}\Sigma x_i \Sigma y_i=\underline{\Sigma x_i y_i-n\overline{xy}}$

$l_{xx}=\Sigma(x_i-\overline{x})^2=\Sigma x_i^2-\frac{1}{n}(\Sigma x_i)^2=\underline{\Sigma x_i^2-n\overline{x}^2}$

$l_{yy}=\Sigma(y_i-\overline{y})^2=\Sigma y_i^2-\frac{1}{n}(\Sigma y_i)^2=\underline{\Sigma y_i^2-n\overline{y}^2}$

可得：

$\hat{\beta}_1=\frac{l_{xy}}{l_{xx}}$

$\hat{\beta}_0=\overline{y}-\hat{\beta}_0\overline{x}$

当$x=\overline{x}$时，$\hat{y}=\overline{y}$ ，即回归直线$\hat{y_i}=\hat{\beta}_0+\hat{\beta}_1 x_i$ 通过点$(\overline{x},\overline{y})$

13.1.3 一元线性回归的检验

$SST=\Sigma(y_i-\overline{y})^2$

离差分解：

1. $y-\overline{y}=(y-\hat{y})+(\hat{y}-\overline{y})$

2. $\Sigma(y_i-\overline{y})^2=\Sigma(y_i-\hat{y}_i)^2+\Sigma(\hat{y}_i-\overline{y})^2$

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

2. 计算检验统计量F $\quad F=\frac{SSR/1}{SSE/(n-2)}=\frac{MSR}{MSE} \quad \sim \quad F(1,n-2)$

3. 作出决策：

   $F>F_\alpha$，拒绝H_0，表明两个变量之间的线性关系是显著的

   $F<F_\alpha$，不拒绝H_0

另一种简便方法：

$S_R=\hat{\beta}^2 l_{xx}=\hat{\beta} l_{xy}$

$S_e=l_{yy}-S_R=l_{yy}-\hat{\beta} l_{xy}$

13.1.4 判定系数

$$R^2=\frac{SSR}{SST}=\frac{\Sigma(\hat{y}-\overline{y})^2}{\Sigma(y_i-\overline{y})^2}=1-\frac{\Sigma(y_i-\hat{y}_i)^2}{\Sigma(y_i-\overline{y})^2}$$

另一种简便方法：

$$R^2=\frac{SSR}{SST}=\frac{S_R}{S_e}=\frac{\hat{\beta}_1 l_{xy}}{\hat{\beta}_1 l_{xy}+l_{yy}-\hat{\beta}_1 l_{xy}}=\hat{\beta}_1\frac{l_{xy}}{l_{yy}}=\frac{l_{xy}^2}{l_{xx}l_{yy}}$$

$$r=\sqrt{R^2}=\sqrt{\hat{\beta}_1\frac{l_{xy}}{l_{yy}}}=\sqrt{\frac{l_{xy}^2}{l_{xx}l_{yy}}}$$

13.2 多元线性回归

13.2.1 多元线性回归模型

多元回归模型与回归方程

描述因变量 y 如何依赖于自变量 $x_1, x_2, ..., x_k$ 和误差项 ε 的方程称为**多元回归模型**。
一般形式： $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$
对误差项 ε 的三个假定

1. 误差项 ε 是一个期望值为 0 的随机变量
2. 对于自变量 $x_1, x_2, x_3, ..., x_k$ 的所有值，误差项 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

多元回归方程：$E(y) = y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$
多元回归方差模型中的基本假定：

1. 自变量$x_1, x_2, ..., x_k$是非随机的、固定的，且相互之间互不相关（无多重共线性），同时样本容量必须大于所要估计的回归系数的个数，即$n > k$；
2. 误差项ε是一个期望值为0的随机变量，即$E(\varepsilon) = 0$；
3. 对于自变量$x_1, x_2, ..., x_k$的所有值，ε的方差σ^2都相同，且无序列相关，即$D(\varepsilon_i) = \sigma^2, \quad Cov(\varepsilon_i, \varepsilon_j) = 0, i \neq j$；
4. 误差项ε是一个服从正态分布的随机变量，即$E(\varepsilon) = 0$。

估计的多元回归方程

用样本统计量 $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$去估计回归方程中的未知参数$\beta_0, \beta_1, \beta_2, ..., \beta_k$ ，就得到估计的多元回归方程
一般形式：
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k$
偏回归系数的定义：$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$
$\hat{\beta}_1$表示：当$x_2, x_3, ..., x_k$不变时，x_1每变动一个单位，因变量y的平均变动量；
$\hat{\beta}_2$表示：当$x_1, x_3, ..., x_k$不变时，x_2每变动一个单位，因变量y的平均变动量；
$\hat{\beta}_3$表示：当$x_1, x_2, , x_4, ..., x_k$不变时，$x_3$每变动一个单位，因变量$y$的平均变动量；
类推...

参数的最小二乘估计

使得残差平方和 $Q = \sum (y_i - \hat{y}_i)^2$ 最小

$$\begin{cases} \left.\frac{\partial Q}{\partial \beta_0}\right|_{\beta_0 = \hat{\beta}_0} = 0 \\ \left.\frac{\partial Q}{\partial \beta_i}\right|_{\beta_i = \hat{\beta}_i} = 0, \quad i = 1, 2, \cdots, k \end{cases}$$

13.2.2 回归方程的拟合优度

多重判定系数

多元回归中也类似于一元线性回归，也有 $SST = SSR + SSE$
多重判定系数 $R^2 = \frac{SSR}{SST}$
多重判定系数是：

1. 多元回归中的回归平方和占总平方的比例，

2. **度量多元回归方程拟合程度**的一个**统计量** ，

3. 反应了因变量 y 中被估计的回归方程所占的比例

计算：

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

解释：

实际意义：在因变量y 取值的变差中，能被因变量y 和自变量$x_1, x_2, ..., x_k$ 的多元回归方程所解释的比例为R^2

例子：

$R^2 = 72.7125\%$

因变量y ： 不良贷款

变量$x_1, x_2, ..., x_k$ ： 比如贷款余额、累计应收贷款、贷款项目个数和固定资产投资额在

不良贷款取值的变差中，能被不良贷款和贷款余额、累计应收贷款、贷款项目个数和固定资产投资额的多元回归方程所解释的比例为75.7125%

⋆在多元统计中，为什么要用**调整的多重判定系数**来比较方程的拟合效果？是如何计算的？

注意：

自变量个数的增加将影响到因变量的变差中**被估计的回归方程**所解释的比例

⇓

增加自变量时，预测的误差变得较小，从而减少残差平方和 SSE

⇓

使得 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

采用**调整的多重判定系数**来避免增加自变量而高估 R^2

$$R_a^2 = 1 - (1 - R^2)(\frac{n-1}{n-k-1})$$

R_a^2 同时考虑了**样本量** n 和模型中**自变量个数** k 的影响

R_a^2 的解释与R^2类似，不同的是，R_a^2同时考虑了因变量n和模型中自变量个数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

因此，在多元回归中，通常采用调整的多重判定系数。

估计标准误差

估计标准误差也是误差项 ε 的方差 σ^2 的一个估计值，在衡量多元回归方程的拟合优度方面有重要作用

$$s_e = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-k-1}} = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{MSE}$$

解释：多元回归中对s_e的解释与一元回归类似

一元回归中对s_e的解释：

1. 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

2. 度量各实际观测点在直线周围的散布状况的一个统计量

含义：根据所建立的多元回归方程，用自变量 $x_1, x_2, ..., x_k$ 来预测因变量 y 时，平均预测误差为 s_e

例子：

$s_e = 1.778752$

自变量 $x_1, x_2, ..., x_k$：贷款余额、累计应收贷款、贷款项目个数和固定资产投资额

因变量 y：不良贷款

s_e 的含义：

根据所建立的多元回归方程，用贷款余额、累计应收贷款、贷款项目个数和固定资产投资额来预测不良贷款时，平均预测误差为1.778752亿元

13.2.3 显著性检验

在多元回归分析中，有两种检验

1. 线性关系检验：检验因变量与**多个**自变量的线性关系是否显著
2. 回归系数检验：对每个回归系数分别进行**单独**的检验，主要用于检验**每个**自变量对因变量的影响是否显著

线性关系检验：在 k 个自变量中，只要有一个自变量与因变量的线性关系显著，F 检验就能通过，但是并不意味着每个自变量和因变量的关系都显著。

回归系数检验：如果某个自变量没有通过检验，就意味着这个自变量对因变量的影响不显著，就没有必要将这个自变量放进回归模型中了。

线性关系检验　F检验

1. 提出假设：$H_0: \beta_1 = \beta_2 = ... = \beta_k = 0 \quad H_1: \beta_1, \beta_2, ..., \beta_k$ 至少有一个不等于0
2. 计算检验统计量 $F = \frac{SSR/k}{SSE/(n-k-1)} \quad \sim \quad F(n, n-k-1)$
3. 做出统计决策，拒绝域 $C = \{F > F_\alpha\}$

回归系数检验　t检验

1. 提出假设：$H_0: \beta_i = 0 \quad H_1: \beta_i \neq 0$
2. 计算检验统计量 $t_i = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}} \quad \sim \quad t(n-k-1)$
3. 做出统计决策，拒绝域 $C = \{|t| > t_{\alpha/2}\}$

$$s_{\hat{\beta}_i} = \frac{s_e}{\sqrt{\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2}}$$

回归系数 β_i 在 $1-\alpha$ 的置信水平下的置信区间为

$$\hat{\beta}_i \pm t_{\alpha/2}(n-k-1)s_{\hat{\beta}_i}$$

13.2.4 多重共线性

多重共线性及其所产生的问题

当模型中有两个或两个以上的自变量批次相关时，则称回归模型中存在**多重共线性**

多重共线性产生的问题：

1. 变量之间高度相关时，会使得回归的结果混乱，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

多重共线性的判别

最简单的一种方法是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

多重共线性的处理

1. 将一个或多个相关的自变量从模型中剔除，使得保留的自变量尽可能不相关
2. 如果要保留所有的自变量：
   - 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得范围内

13.2.5 利用回归方程进行预测

借助于计算机

13.2.6 变量选择与逐步回归

变量选择过程

变量选择的意义：避免多重共线性，在建立模型之前就对收集到的自变量进行检测，去掉不必要的自变量，使得模型变得更容易，更具有可操作性，更容易解释

选择自变量的原则：对统计量进行显著性检验

显著性检验的依据：将一个或多个自变量引入回归模型中时，是否使得残差平方和 SSE 显著减少；如果显著减少则有必要将这个自变量引入回归模型

向前选择

向前选择的原理：从模型中没有自变量开始

1. 先对 k 个自变量 $x_1, x_2, ..., x_k$ 分别拟合于因变量 y 的**一元回归模型**，然后找出 F 统计量的值最大的模型及其自变量 x_i，将其首先引入模型
2. 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 $x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_k$ 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，分别考虑这 $k-1$ 个线性模型，挑选出 F 统计量最大的含有两个自变量的模型，将 F 统计量最大的那个自变量 x_j 引入模型
3. .....
4. 重复上述过程，直到没有显著的或者所有的都引入了。

向前选择变量的方法就是不断的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（ F 检验）为止。

向后剔除

向后剔除的原理：

1. 先对因变量拟合包括所有 k 个自变量的线性回归模型；
2. 考虑 $p(p < k)$ 个去掉一个自变量的模型（每一个模型都有 $k-1$ 个自变量），使得模型的 SSE 值减小最小的自变量被挑选出来并从模型中剔除；
3. 考察 $p-1$ 个再去掉一个自变量的模型（每一个模型都有 $k-2$ 个自变量），使得模型的 SSE 值减小最小的自变量被挑选出来并从模型中剔除；
4. .....
5. 重复上述过程，直到剔除一个自变量不会使得 SSE 显著减小为止。

逐步回归

逐步回归的原理和应用:

1. 前两步与向前选择法相同
2. 在增加一个自变量之后，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
3. 不停的增加变量并考虑剔除以前增加的变量的可能性，直到增加变量不会导致 SSE 显著减少

在逐步回归法中，前面步骤中增加的自变量可能在后面的步骤中剔除，前面步骤中剔除的自变量可能在后面的步骤中重新进入模型

Chapter 14

多元线性回归

14.1 多元回归模型

14.1.1 多元回归模型与多元回归方程

多元回归模型 描述因变量y如何依赖于自变量$x_1, x_2, ..., x_k$和误差项ε的方程称为多元回归模型。

一般形式：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

多元回归方程 描述因变量y的期望值与自变量$x_1, x_2, ..., x_k$之间的关系

公式：

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

多元线性回归模型中的基本假定 a

对自变量$x_1, x_2, ..., x_k$的假定：

自变量$x_1, x_2, ..., x_k$是非随机的、固定的，且相互之间互不相关（无多重共线性）；

同时样本容量必须大于所要估计的回归系数的个数，即$n > k$；

对误差项ε的假定：

误差项ε是一个期望值为0的随机变量，即$E(\varepsilon) = 0$；

误差项ε是一个服从正态分布的随机变量，即$E(\varepsilon) = 0$；

对于自变量$x_1, x_2, ..., x_k$的所有值，ε的方差σ^2都相同，且无序列相关，即 $D(\varepsilon_i) = \sigma^2$， $Cov(\varepsilon_i, \varepsilon_j) = 0, i \neq j$。

期望为0，正态分布，方差相同

估计的多元回归方程 用样本统计量$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$去估计回归方程中的未知参数$\beta_0, \beta_1, \beta_2, ..., \beta_k$，就得到估计的多元回归方程

一般形式： $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k$

偏回归系数的解释 偏回归系数的定义：$\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$

$\hat{\beta}_1$表示：当$x_2, x_3, ..., x_k$不变时，x_1每变动一个单位因变量y的平均变动量；

$\hat{\beta}_2$表示：当$x_1, x_3, ..., x_k$不变时，x_2每变动一个单位因变量y的平均变动量；

$\hat{\beta}_3$表示：当$x_1, x_2, , x_4, ..., x_k$不变时，$x_3$每变动一个单位因变量$y$的平均变动量；

类推...

注意： $\hat{\beta}_0$不是偏回归系数

14.1.2 参数的最小二乘估计

利用Excel进行回归

14.1.3 多重判定系数

概念:多元回归中也类似于一元线性回归，也有 $SST = SSR + SSE$

多重判定系数

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

多重判定系数是：多元回归中的回归平方和占总平方和的比例，度量多元回归方程拟合程度的一个统计量，反应了因变量 y 中被估计的回归方程所占的比例

多重判定系数的计算和解释 计算： $R^2 = \frac{SSR}{SST}$

解释：

实际意义：在因变量y 取值的变差中，能被因变量y 和自变量$x_1, x_2, ..., x_k$ 的多元回归方程所解释的比例为R^2

例子： $R^2 = 72.7125\%$

因变量y ： 不良贷款

自变量$x_1, x_2, ..., x_k$ ： 比如贷款余额、累计应收贷款、贷款项目个数和固定资产投资额

在不良贷款取值的变差中，能被不良贷款和贷款余额、累计应收贷款、贷款项目个数和固定资产投资额的多元回归方程所解释的比例为75.7125

14.1.4 调整的多重判定系数的概念

为什么在多元统计中要用调整的多重判定系数来比较方程的拟合效果？是如何计算的？

答：自变量个数的增加将影响到因变量的变差中被估计的回归方程所解释的比例增加自变量时，预测的误差变得较小，从而减少残差平方和 SSE 使得 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

自变量个数增多 $\Rightarrow$ 预测的误差变小 $\Rightarrow$ 残差平方和减小 $\Rightarrow$ R^2变大

采用调整的多重判定系数来避免增加自变量而高估 R^2

$$R_a^2 = 1 - (1 - R^2)(\frac{n-1}{n-k-1})$$

R_a^2 同时考虑了样本量 n 和模型中自变量个数 k 的影响

调整的多重判定系数的计算和解释 计算： $R_a^2 = 1 - (1 - R^2)(\frac{n-1}{n-k-1})$

解释：copy上面的解释，实际上差不多，只是克服了增加自变量个数造成的影响。

实际意义：在因变量y 取值的变差中，能被因变量y 和自变量$x_1, x_2, ..., x_k$ 的多元回归方程所解释的比例为R^2

例子： $R^2 = 72.7125\%$

因变量y ： 不良贷款

自变量$x_1, x_2, ..., x_k$ ： 比如贷款余额、累计应收贷款、贷款项目个数和固定资产投资额

在不良贷款取值的变差中，能被不良贷款和贷款余额、累计应收贷款、贷款项目个数和固定资产投资额的多元回归方程所解释的比例为75.7125%

R_a^2的解释与R^2类似，不同的是，R_a^2同时考虑了因变量n和模型中自变量个数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 因此，在多元回归中，通常采用调整的多重判定系数

14.1.5 估计标准误差的计算和解释

估计标准误差也是误差项 ε 的方差 σ^2 的一个估计值，在衡量多元回归方程的拟合优度方面有重要作用

计算：

$$s_e = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-k-1}} = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{MSE}$$

解释：多元回归中对s_e的解释与一元回归类似

一元回归中对s_e的解释：反映了用估计的回归方程预测因变量y时预测误差的大小度量各实际观测点在直线周围的散布状况的一个统计量

含义：根据所建立的多元回归方程，用自变量 $x_1, x_2, ..., x_k$ 来预测因变量 y 时，平均预测误差为 s_e

例子： $s_e = 1.778752$

自变量 $x_1, x_2, ..., x_k$：贷款余额、累计应收贷款、贷款项目个数和固定资产投资额

因变量 y：不良贷款

s_e 的含义：根据所建立的多元回归方程，用贷款余额、累计应收贷款、贷款项目个数和固定资产投资额来预测不良贷款时，平均预测误差为1.778752亿元

14.2 显著性检验

14.2.1 在多元回归中，显著性检验可以分为：

线性关系检验：检验因变量与多个自变量的线性关系是否显著

回归系数检验：对每个回归系数分别进行单独的检验，主要用于检验每个自变量对因变量的影响是否显著

14.2.2 线性关系检验 F检验

在 k 个自变量中，只要有一个自变量与因变量的线性关系显著， F 检验就能通过，但是并不意味着每个自变量和因变量的关系都显著。

14.2.3 回归系数检验 t检验

如果某个自变量没有通过检验，就意味着这个自变量对因变量的影响不显著，就没有必要将这个自变量放进回归模型中了。

线性关系检验的步骤： 检验线性关系

提出假设： $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0 \quad H_1 : \beta_1, \beta_2, ..., \beta_k$ 至少有一个不等于0

计算检验统计量

$$F = \frac{SSR/k}{SSE/(n-k-1)} \quad \sim \quad F(n, n-k-1)$$

做出统计决策，拒绝域 $C = \{F > F_\alpha\}$

回归系数检验的步骤： 检验偏回归系数

提出假设： $H_0 : \beta_i = 0 \quad H_1 : \beta_i \neq 0$

计算检验统计量 $t_i = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}} \quad \sim \quad t(n-k-1)$

做出统计决策，拒绝域 $C = \{|t| > t_{\alpha/2}\}$

$$s_{\hat{\beta}_i} = \frac{s_e}{\sqrt{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}}$$

$s_{\hat{\beta}_i}$是回归系数$\hat{\beta}_i$ 的抽样分布的标准差

回归系数 β_i 在 $1-\alpha$ 的置信水平下的置信区间为

$$\hat{\beta}_i \pm t_{\alpha/2}(n-k-1) s_{\hat{\beta}_i}$$

14.3 多重共线性

多重共线性的概念 当模型中有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

多重共线性产生的问题 变量之间高度相关时，会使得回归的结果混乱，甚至会把分析引入歧途多重共线性可能对参数估计值的正负号产生影响

多重共线性的判别 最简单的一种方法是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计算其相关系数并对其进行检验

多重共线性的处理 将一个或多个相关的自变量从模型中剔除，使得保留的自变量尽可能不相关如果要保留所有的自变量：避免根据 t 统计量对单个参数进行检验对因变量 y 值的推断（估计或预测）限定在自变量样本值得范围内

14.4 变量选择与逐步回归

向前选择的原理 不断的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（ F 检验）为止。

先对 k 个自变量 $x_1, x_2, ..., x_k$ 分别拟合于因变量 y 的一元回归模型，

然后找出 F 统计量的值最大的模型及其自变量 x_i ，将其首先引入模型在已经引入模型的 x_i 的基础上，

再分别拟合 $k-1$ 个自变量 $x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_k$ 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，分别考虑这 $k-1$ 个线性模型，挑选出 F 统计量最大的含有两个自变量的模型，将 F 统计量最大的那个自变量 x_j 引入模型

.....

重复上述过程，直到没有显著的或者所有的都引入了。

向后剔除的原理 先对因变量拟合包括所有 k 个自变量的线性回归模型；

考虑 $p(p < k)$ 个去掉一个自变量的模型（每一个模型都有 $k-1$ 个自变量），使得模型的 SSE 值减小最小的自变量被挑选出来并从模型中剔除；

考察 $p-1$ 个再去掉一个自变量的模型（每一个模型都有 $k-2$ 个自变量），使得模型的 SSE 值减小最小的自变量被挑选出来并从模型中剔除；

.....

重复上述过程，直到剔除一个自变量不会使得 SSE 显著减小为止。

逐步回归的原理 前两步与向前选择法相同

在增加一个自变量之后，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

不停的增加变量并考虑剔除以前增加的变量的可能性，直到增加变量不会导致 SSE显著减少

在逐步回归法中，前面步骤中增加的自变量可能在后面的步骤中剔除，前面步骤中剔除的自变量可能在后面的步骤中重新进入模型

Chapter 15

多元统计分析

寻找多个变量的代表：主成分分析、因子分析

把对象分类：聚类分析

两组变量之间的相关：典型相关分析

列联表行变量和列变量之间的关系：对应分析

15.1 多元正态分布

- 多元正态分布的基本概念

  随机向量、分布函数与密度函数、多元变量的独立性、随机向量的数字特征（均值向量、协方差阵、相关阵）

- 统计距离

  欧氏距离、统计距离、马氏距离

- 多元正态分布

  定义、性质、条件分布和独立性、

- 均值向量和协方差阵的估计

- 常用分布及抽样分布

  χ^2 分布与 $Wishart$ 分布、 t 分布与 T^2 分布、中心 F 分布与 $Wilks$ 分布

课后思考：

1. 在数据处理时，为什么通常要进行标准化处理？
2. 欧氏距离和马氏距离的优缺点是什么？

15.2 均值向量和协方差阵的检验

1. 均值向量的检验
2. 协方差阵的检验

15.3 聚类分析

聚类分析将个体或对象分类，使得同一类中的对象之间的相似性比其他类的对象的相似性更强。
目的：使得类内对象的同质性最大化和类于类间对象的异质性最大化、把相似的研究对象归为一类
多元分析三大方法：聚类分析、回归分析、判别分析

1. 聚类分析的基本思想
   根据一批样本的多个观测指标，具体找出一些能够度量样品或指标之间相似程度的统计量，以这些统计量作为划分类型的依据，把一些相似程度比较大的样本（或指标）聚合为一类，把另外一些相似程度较大的样本（或指标）合并为另一类
   相似样品（或指标）的集合称为类

2. 聚类分析的分类方法：系统聚类法、模糊聚类法、K均值法、有序样品的聚类、分解法、加入法

3. 相似性度量
   对样品进行聚类时，用距离来刻画
   对指标进行聚类时，用相关系数或某种关联性度量

   距离：绝对值距离、欧氏距离、Minkowski距离、切比雪夫距离、兰氏距离、马氏距离...
   相似系数：夹角余弦、相关系数

4. 类和类的特征
   类于类的距离：最短距离法、最长距离法、重心法、类平均法、离差平方和法

5. 系统聚类法
   开始时，有多少点就是多少类 → 把最近的两个点分为一类 → 依此类推 → 只剩一个类

6. 模糊聚类分析

7. K-均值聚类和有序样品的聚类

问题:

1. 什么是样品

2. 什么是指标（即变量）

这两种方法在数学上是对称的
R型聚类：按照观测值对变量进行分类
Q型聚类：按照变量对观测值进行分类

如何度量距离远近:
点间距离：欧式距离、平方欧式距离、绝对距离、Chebychev距离、Minkovski距离、夹角余弦、Pearson相关系数
类间距离：最短距离法、最长距离法、重心法、类平均法、离差平方和法

事先确定要分多少类：k均值聚类
事先不同确定分多少类：分层聚类（也叫系统聚类）

1. 开始时，有多少点就是多少类

2. 把最近的两个点分为一类

3. 依此类推

4. 只剩一个类

聚类结果受所选择的原始变量的影响，选择的变量不同，结果不同

聚类的目的是使得各类之间的距离尽可能远，而类中点此间的距离尽可能近

最短距离法的缺点：链接聚合。类于类的距离为所有距离中的最短者，两类合并之后，它与其余类的距离缩小了，这样容易形成一个比较大的类，在树状图中会有一个延申的链状结构。

最长距离法克服了最短距离法连接聚合的缺陷

15.4 判别分析

1. 判别分析的基本思想
   判别分析适用于被解释变量是非度量变量的情形，主要目的是识别一个个体所属类别
   判别分析的假设之一：每一个判别变量（解释变量）不能是其他判别变量的线性组合
   判别分析的假设之二：各组变量的协方差相等
   判别分析的假设之三：各判别变量遵从正态分布

2. 距离判别

3. 贝叶斯判别
   贝叶斯思想：假定对研究对象有了一定的认识，常用先验概率分布来描述这种认识，然后取得一个样本，用样本来修正已有的认识（先验概率分布），得到后验概率分布，各种统计推断都通过**后验概率分布**来进行

$$P(B_i|A) = \frac{P(B_j)P(A|B_j)}{\sum_{j=1}^{n} P(B_j)P(A|B_j)}$$

4. Fisher判别
   思想：投影，将 k 组 p 维数据投影到某一个方向，使得组与组之间的投影尽可能分开。

5. 逐步判别
   凡具有筛选变量能力的判别方法统称为逐步判别法

15.5 主成分分析

利用降维的思想，在损失信息很少的前提下，把多个指标转化为几个综合指标

在保留原始数据尽可能多的情况下达到降维的目的，从而简化问题，抓住问题的主要矛盾

相关矩阵是将原始数据标准化后的协方差矩阵

主成分分析不要求数据来自正态总体

1. 主成分分析的基本原理
   研究如何通过原来变量的少数几个线性组合来解释原来变量绝大多数信息
   二维空间的几何意义：坐标轴转换，各坐标轴的方向就是原始数据变差最大的方向

2. 总体主成分及其性质

3. 样本主成分的导出

4. 主成分分析步骤

   - 根据研究问题选择初始变量
   - 根据初始变量特性判断由协方差矩阵求主成分还是由相关阵求主成分

- 求协方差阵或相关阵的特征根与相应标准特征向量
- 判断是否存在明显的**多重共线性**，若存在，则回到第一步
- 得到主成分的表达式，并确定主成分个数，选取主成分

主成分分析和因子分析：把变量维数降低以便于描述、理解和分析
主成分分析和因子分析只能对互相相关的数量变量进行降维
如果变量没有近似的多维正态分布，降维可能不理想
依赖于原始变量
不能把主成分分析的特征向量按照特征值大小的加权平均来得到“综合指数”

1. 结果太多
2. 正交

主成分：互相正交的新变量是原来变量的线性组合
目的：用少数的几个互相正交的向量来代表原始数据中较多的相关的变量

- 从原理上是寻找椭球的所有主轴，因此，原来有多少个变量，就有多少个主成分
- 选择的标准：
  - 被选的主成分所代表的主轴的长度之和占了主轴长度之和的大部分
  - 各个成分的方差
    主成分为数据相关阵的特征向量，每个成分的方差为相应的特征值
- 悬崖碎石图的原则：如果该图不陡，那么主成分分析的结果一定不好
  不要为了凑够百分数而选取悬崖下面的“碎石”
- 载荷表
- 主成分载荷：主成分和原来各变量的线性相关系数，其绝对值越大，主成分对该变量的代表性就越大
- 载荷图：第一主成分是横坐标，第二主成分是纵坐标
- 把因子（成分）写在方程的左边，把原变量写在方程的右边
  $y_1 = 0.80x_1 - 0.72x_2 + ... + 0.92x_{10}$

15.6 因子分析

因子分析是主成分分析的推广

1. 因子分析的基本理论
   基本思想：依据相关性大小将原始变量分组，使得同族内的变量之间的相关性较高，而不同组的变量间的相关性则比较低
   因子分析还可用于对变量或样品的分类处理
   R型因子分析：研究变量之间的相关关系
   Q型因子分析：研究样品之间的相关关系
2. 因子载荷的求解
3. 因子分析的步骤与逻辑框图

- 主成分分析从原理上是寻找椭球的所有主轴，因此，原来有多少个变量，就有多少个主成分

- 因子分析是实现确定要找多少个成分
- 碎石图
- 因子载荷
- 与主成分分析不同，把因子（成分）写在方程的右边，把原变量写在方程的左边

  $x_1 = -0.6604421y_1 - 0.3320026y_2$

15.7 对应分析

对应分析是R型因子分析和Q型因子分析的结合

1. 列联表及列联表分析
2. 对应分析的基本理论
3. 对应分析的步骤及逻辑框图

因子分析：或者对变量进行分析，或者对样品进行分析，而且常常把每一种分析结果画出载荷图来看各个变量之间的接近程度
典型相关分析：只研究列中两组变量之间的关系
对应分析：若干行变量与若干列变量之间的关系，或者是列联表中行变量与列变量的各个水平之间的相互关系

对应分析把一对列向量和一对行向量同时反映到同一张图上。

15.8 典型相关分析

1. 典型相关分析的基本理论及方法
2. 典型相关分析的步骤及逻辑框图

两组变量之间的相关：典型相关分析
找到这两组变量之间的线性组合的**系数**使得这两个由线性组合生成的变量（和其他线性组合相比）之间的**相关系数**最大

思想：将每一组变量用一个变量来代表，那么，**两组**变量之间的相关就可以转化为**两个**变量之间的相关。但是要注意，这两个变量之间一定要互相联系，主成分分析找出的两组变量不一定能互相联系。

目的：在两组变量之间各找到一个或多个有综合意义的代表变量，通过研究这两个代表变量之间的关系来考察两组变量之间的关系。

典型相关系数