1 统计学基础

1.1 统计学

- 什么是**统计学**？统计学是寻找更好的数据应用方法的学科。
- 为了整理、归纳现有数据而产生的统计学分支，叫作**描述统计**。
- 为了估计不在我们手中的未知数据而产生的统计学分支叫作**统计推断**。
- **使用现有数据能推断未知数据**——这可以说是学习统计学给我们带来的最大好处。
- **样本**是指现有数据。
- **总体**是指既包含现有数据也包含未知数据的全部数据。
- **只使用样本这一部分数据来讨论总体这一全部数据**就是统计推断的目标。

1.2 获取样本的过程

- 根据随机法则变化的量叫作**随机变量**。
- 由随机变量得来的具体数值叫作**样本值**。
- 从总体中获取样本叫作**抽样**。
- 随机选择总体中各个元素的方法叫作**简单随机抽样（随机抽样）**。
- 样本的大小或现有数据的个数叫作**样本容量**。
- 调查完整的总体叫作**普查**。
- 只调查总体的一部分叫作**抽样调查**。

1.3 抽样过程的抽象描述

- **概率**使用它的英文$probability$的首字母P来表示。获得某个数据的概率记为$P(\text{数据})$。
- **概率分布（分布）**表示随机变量及其概率之间的关系。
- 当某数据和某种概率分布相符时，就叫作**服从概率分布**。
- 总体服从的概率分布叫作**总体分布**。
- 像投掷骰子的实验这种结果范围无穷大的总体叫作**无限总体**。
- 使用瓮中取球的实验来描述多种现象的模型叫作**瓮模型**。

1.4 描述统计基础

- “定量”的意思就是**数值之间的差距所代表的意义是等价的**。
- 能够定量表示的数据叫作**定量变量（数值变量）**。
- 定量变量又分为两种。1条、2条这种只取整数的数据叫作**离散变量**。2.3cm、4.25cm这种会取到小数点之后的值且变化连续的数据叫作**连续变量**。
- 不能定量表示的数据叫作**分类变量**。
- 如果把鱼分为青鳉和鲤鱼2种，则“青鳉”等种类名称就是分类变量，这也叫作**名义尺度**。有的分类变量叫作**顺序尺度**，比如，把鱼按大、中、小进行区分， 像这种有顺序的分类就是顺序尺度。
- 在定量变量中，我们经常会看到数值被分成几个范围，这些范围就叫作**组**。
- 代表组的值叫作**组中值**，取组的最大值和最小值之间的中间数值。
- **频数**就是某个数据出现的次数。
- **频数分布**是每个组中数据的频数的排列。
- 频数占总数的比例叫作**频率**。

- 把组按从小到大的顺序排列，将频数相加，得到的和叫作**累积频数**。
- 按同样方法得到的频率的和叫作**累积频率**。
- 表示频数分布的图叫作**直方图**。在直方图中，横轴是组，纵轴是频数。
- 用于统计数据的值叫作**统计量**。
- 最常用的统计量是**均值**。
- 需要注意的是，均值被用作代表已知数据（样本）的值，即**代表值**。
- 在统计学中，均值经常被称为期望值。我们可以把期望值理解为能够用于未知数据的均值。
- **方差**用来表示数据与均值（期望值）之间相差多少。

1.5 总体分布的推断

- 统计学中会通过**做假设**来简化计算。
- 在统计学中也会为总体分布做假设，**正态分布**就是其中一种，既容易计算又符合数据。

1.6 概率质量函数与概率密度函数

- 在将数据作为参数时，所得函数值是概率的函数叫作**概率质量函数**。
- 离散变量可以直接计算概率，而连续变量不能直接计算概率，于是人们就用**概率密度**来代替概率。我们可以 把概率密度看作用于连续变量的类似于概率的概念。
- 可以认为**积分是加法的扩展**。
- 在将数据作为参数时，所得函数值是概率密度的函数叫作**概率密度函数**。
- **参数**是用于定义概率分布的值。
- 正态分布有两个参数，分别是均值（期望值）μ和方差σ^2。若将随机变量记为x，则正态分布的概率密度函数就记为$\mathcal{N}(x)$。通过计算$\mathcal{N}(x)$，可以求得随机变量对应的概率密度。有时也将正态分布的概率密度函数记为$\mathcal{N}(x|\mu,\sigma^2)$，以明确地表示参数。
- **估计参数最直接的思路就是把样本的统计量看作总体分布的参数**。
- 但是样本的统计量和参数之间普遍存在差别，我们必须认识到估计出来的参数存在**估计误差**。
- 在进行参数估计时，如果要考虑估计误差，可以使用**区间估计**等方法。
- 在存在估计误差的情况下依然要证明某个想法时，可以使用**假设检验**。

1.7 统计量的计算

- 样本的均值μ可通过该式求出：$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$，准确来说，这个均值叫作**算术平均值**。
- 离散变量的期望值μ可以通过求“取值的概率$\times$取到的值”的总和计算出来：$\mu = \sum_{i=1}^{N} P(x_i)\cdot x_i$。
- 总体的均值叫作**总体均值**。样本的均值叫作**样本均值**。总体均值与样本均值经常存在差距，但不会偏离。
- 方差用来表示数据与均值（期望值）之间相差多少，由该式求得：$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$。其中，$x_i-\mu$叫作**偏差**。各个偏差的平方的总和，即$\sum(x_i-\mu)^2$，叫作**偏差平方和**。
- 使用概率$P(x_i)$，可以将方差表示为：$\sigma^2 = \sum_{i=1}^{N} P(x_i)\cdot(x_i-\mu)^2$。数据和期望值之间相差越大，$(x_i-\mu)^2$的值越大。$(x_i-\mu)^2$可用来表示数据与期望值之间的距离。因此，方差也就是数据与期望值之间的距离的期望值。准确来说，此处的方差叫作**样本方差**。
- 样本方差和总体方差之间存在偏离，前者比后者偏小。**无偏方差**就是为了修正这个偏离而出现的。通过该式计算无偏方差：$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu)^2$。
- 标准差通过对方差取平方根而得出：$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu)^2}$。这里的方差一般使用无偏方差。

1.8 概率论基础

- **集合**是由客观标准定义的事物一起形成的总体。
- 设有集合A，某个事物a是A的**元素**，则记作$a \in A$，读作a**属于**A。
- 设有两个集合A、B。如果当$a \in A$时，$a \in B$，则称A是B的**子集**，记作$A \subset B$。
- 在比较集合时，人们经常使用**维恩图**。
- 对于A、B两个集合，**交集**$A \cap B$的定义为：$A \cap B = \{a; a \in A且a \in B\}$。
- 对于A、B两个集合，**并集**$A \cup B$的定义为：$A \cup B = \{a; a \in A或a \in B\}$。
- 对于A、B两个集合，**差集**$A - B$的定义为：$A - B = \{a; a \in A且a \notin B\}$。
- 不含任何元素的集合叫作**空集**，记作$\emptyset$。
- 设有集合S，当所研究的问题只考虑S的子集时，称S为**全集**。
- 已知全集S，关于S的子集A有如下关系成立，则称集合A^c为A的**补集**。
- 可能发生的实验结果称为**样本点（ω）**，样本点的总体的集合称为**样本空间（Ω）**。
- 样本空间的子集叫作**事件**。与集合一样，事件也有**并事件**与**交事件**的定义。
- 只由一个样本点组成且不可再分解的事件叫作**基本事件**。含有多个样本点且可以分解为多个基本事件的事件叫作**复合事件**。与空集类似，不含任何样本点的事件叫作**空事件**。
- 当$A \cap B = \emptyset$时，或者当事件之间没有重叠时，称事件A和B是**互斥事件**。
- **概率的加法公式**是，如果A、B是互斥事件，则满足：$P(A \cup B) = P(A) + P(B)$。去掉互斥的条件，将概率的加法公式一般化：$P(A \cup B) = P(A) + P(B) - P(A \cap B)$。
- 以发生事件B为前提条件，事件A发生的概率叫作**条件概率**，记作$P(A|B)$，定义为：$P(A|B) = \frac{P(A \cap B)}{P(B)}$。
- 将条件概率的公式进行变形，可得到：$P(A \cap B) = P(B) \cdot P(A|B)$。该式称为**概率的乘法公式**。
- **独立事件**：当$P(A \cap B) = P(A) \cdot P(B)$成立时，称事件$A$、$B$相独立。变形后就是$P(A|B) = P(A)$。

1.9 随机变量与概率分布

- 离散型数据可以通过求概率的总和来得到各种事件的概率，而连续型数据则要通过求概率密度的积分来得到事件的概率。二者的方法是互相对应的。
- 积分就是曲线下方的面积大小。当$n \to \infty$时，即**当区间分成无穷多个时，面积之和就叫作积分**，表示为$\lim_{n\to\infty}[\sum_{i=1}^{n} f(x_i) \times \Delta x] = \int_a^b f(x)dx$。
- 正态分布的概率密度函数为：$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{[-\frac{(x-\mu)^2}{2\sigma^2}]}$。
- **独立同分布**即随机变量服从同一分布且相互独立。

3 使用$Python$进行数据分析

3.1 使用$Python$进行描述统计：单变量

- **单变量**数据是指只有一种类型的数据，例如鱼的体长。
- **标准化**就是把均值转化为0，把标准差（方差）转化为1。要使得均值为0，只需用所有样本减去均值即可。同样，要使得数据的标准差（方差）为1，只需用所有样本除以标准差即可。
- 把数据按升序排列，位置在最中间的数就是**中位数**。
- 均值易受**极端值**的影响，而中位数不易受极端值的影响。因此，**中位数对于极端值更有稳健性**。

- 把数据按升序排列，处在25%和75%位置的数就是**四分位数**。

3.2 使用$Python$进行描述统计：多变量

- 研究两个连续变量之间的关系时使用的统计量叫作**协方差**。变量x、y的协方差$Cov(x,y)$的计算公式为：$Cov(x,y)=\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)$。其中，$\mu_x$、$\mu_y$分别是变量$x$、$y$的均值，$N$是样本容量。
- 和方差一样，协方差的公式中也可以使用$N-1$作为分母：$Cov(x,y)=\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)$。
- 把多个变量的方差和协方差放在一起形成的矩阵，叫作**协方差矩阵**。变量x、y的协方差矩阵为：$Cov(x,y)=\begin{bmatrix}\sigma_x^2 & Cov(x,y)\\ Cov(x,y) & \sigma_y^2\end{bmatrix}$。其中，$\sigma_x^2$、$\sigma_y^2$分别是$x$、$y$的方差。
- 对于变量x、y，该式计算所得的ρ_{xy}叫作**皮尔逊积矩相关系数（相关系数）**：$\rho_{xy}=\frac{Cov(x,y)}{\sqrt{\sigma_x^2\sigma_y^2}}$。该式也可以看作将协方差标准化成最大值为1、最小值为-1而得出的。
- 把多个变量的相关系数放在一起得到的矩阵，叫作**相关矩阵**。变量x、y的相关矩阵为：$Cov(x,y)=\begin{bmatrix}1 & \rho_{xy}\\ \rho_{xy} & 1\end{bmatrix}$。

3.4 用$Python$模拟抽样

- 随机得到的数叫作**随机数**。有些领域会将其看作随机变量。
- 把抽出的样本放回总体再重新抽样叫作**放回抽样**，抽出的样本不放回总体的抽样叫作**不放回抽样**。

3.5 样本统计量的性质

- 试验可以在完全相同的条件下进行多次，这叫作**重复试验**。
- 在能够重复试验的前提下重复进行试验的次数叫作**试验次数**。
- **样本分布**是样本的统计量所服从的概率分布。
- 样本均值的标准差的理论值叫**标准误差**，计算公式：$标准误差=\frac{\sigma}{\sqrt{N}}$，其中，$\sigma$是标准差，$N$是样本容量。样本容量越大，标准误差就越小。
- 估计量的期望值相当于真正的参数的特性叫作**无偏性**。
- 样本容量越大，估计量越接近真正的参数的特性称为**一致性**。
- 所谓**大数定律**，就是样本容量越大，样本均值越接近总体均值。
- 对于任意总体分布，样本容量越大，随机变量的和的分布越接近正态分布，这就是**中心极限定理**。

3.6 正态分布及其应用

- 在该式中：$F(X)=P(X\leq x)$，对于随机变量X，当x为实数时，$F(X)$叫作**累积分布函数（分布函数）**。简单来说，累积分布函数可以计算随机变量小于等于某个值的概率。
- 数据小于等于某个值的概率叫作**左侧概率**。借助累积分布函数可以得到左侧概率。
- 与上述概念相反，能得到某个概率的那个值叫作**百分位数**，也叫左侧百分位数。
- 均值为0、方差（或标准差）为1的正态分布叫作标准正态分布，即$\mathcal{N}(x|0,1)$。
- 统计量t值的计算方法为：$t=\frac{\hat{\mu}-\mu}{\hat{\sigma}/\sqrt{N}}$。其中，$\hat{\mu}$为样本均值，$\mu$为总体均值，$\hat{\sigma}$为实际样本的无偏标

准差（无偏方差的平方根），N为样本容量。文字描述为：t值 $=\frac{\text{样本均值}-\text{总体均值}}{\text{标准误差}}$。

- t值的公式与标准化公式类似。标准化就是把均值转化为0，把方差转化为1。t值可以理解为对样本均值进行标准化，然而这个计算的除数不是标准误差的理论值，它来自实际样本，不能把方差转化为1。
- 当总体服从正态分布时，t值的样本分布就是t分布。
- 设样本容量为N，$N-1$就叫作自由度。t分布的均值为0，方差稍大于1。设自由度为n（n大于2），则t分布的方差为：$t(n)$的方差 $=\frac{n}{n-2}$。自由度（或样本容量）越大，方差越接近1，t分布越接近标准正态分布；样本容量越小，t分布越远离标准正态分布。
- t分布的意义就是**在总体方差未知时也可以研究样本均值的分布**。

3.7 参数估计

- 直接指定总体分布的参数为某一值的估计方法叫作**点估计**。
- 估计值具有一定范围的估计方法叫作**区间估计**。我们使用概率的方法计算这个范围。
- **置信水平**是表示区间估计的区间可信度的概率。
- 满足某个置信水平的区间叫作**置信区间**。
- 置信区间的下界值与上界值叫作**置信界限**，这两个数值分别叫作**下置信界限**与**上置信界限**。

3.8 假设检验

- 通过样本对总体进行统计学上的判断叫作**假设检验（检验）**，其特征是使用概率论的语言来描述判断。
- **单样本t检验**，研究对象：均值，研究目标：均值是否与某个值存在差异。
- **显著性差**：具有显著性的差异。
- **均值差异大不代表存在显著性差异**，在研究显著性差异时，必须考虑样本容量和方差。
- 一开始提出来并用于拒绝的对象叫作**零假设**。
- 和零假设对立的假设叫作**备择假设**。
- 样本与零假设之间的矛盾的指标就是**p值**。p值越小，零假设和样本之间越矛盾。
- 拒绝零假设的标准叫作**显著性水平（危险率）**。当p值小于显著性水平时，拒绝零假设。显著性水平多使用5%这个数。
- 检验薯片均重是否小于50g叫作**单侧检验**，检验薯片均重是否和50g存在差异叫作**双侧检验**。

3.9 均值差的检验

- **双样本t检验**，判断2种变量的均值是否有差异。
- **配对样本t检验**用于研究在两个不同条件下对同一对象进行测量所得到的值的区别，比如服药前后的体温变化。通过单样本t检验观察体温变化的均值是否与0存在差异。
- **独立样本t检验**的关注重点是两组数据均值的差，而配对样本t检验则是先求数据的差值再进行单样本t检验。
- 独立样本t检验的t值计算公式为：$t=\frac{\hat{\mu}_x-\hat{\mu}_y}{\sqrt{\hat{\sigma}_x^2/m+\hat{\sigma}_y^2/n}}$。其中，$\hat{\mu}_x$为$x$的样本均值，$\hat{\mu}_y$为$y$的样本均值，$m$为$x$的样本容量，$n$为$y$的样本容量，$\hat{\sigma}_x^2$为$x$的无偏方差，$\hat{\sigma}_y^2$为$y$的无偏方差。
- 有些传统教材指出，要先检查数据的同方差性，再进行独立样本t检验。
- 任意改变 p 值的行为就叫作**p值操纵**。
- **作为一名数据分析师，唯有数据是最应该诚恳对待的。**

3.10 列联表检验

- 实际得到的观测数据叫作**观测频数**。
- 如果按钮的颜色对吸引力完全没有影响，我们得到的数据叫作**期望频数**。
- χ^2**统计量**：$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(Q_{ij}-E_{ij})^2}{E_{ij}}$，其中，$Q_{ij}$是第$i$行第$j$列的观测频数，$E_{ij}$是第$i$行第$j$列的期望频数。
- 2行2列的表格对应的χ^2统计量的样本分布被证明近似于自由度为1的χ^2分布。
- 注意，χ^2检验要求所有期望频数不小于5。

3.11 检验结果的解读

- 零假设正确却拒绝了零假设的行为叫作**第一类错误**。
- 反过来，零假设错误却接受了零假设的行为叫作**第二类错误**。
- 第一类错误的概率是可控的，第二类错误的概率是不可控的，这叫作假设检验的**非对称性**。
- **假设检验无法确定第二类错误的概率**。

4 统计模型基础

4.1 统计模型

- **模型**是现实世界的抽象。
- **建模**就是制作模型，统计建模就是制作统计模型。
- 使用与现实世界对应的模型**有助于我们理解和预测现实事物**。
- 我们要制作的就是**面向复杂世界的简单模型**。
- **观察的切入点和建立的模型都可以随着分析目的的改变而变化**。
- **数学模型**使用数学式来表示现象。
- **概率模型**是数学模型中用概率的语言描述的模型。
- **统计模型**是基于数据建立的概率模型。
- 统计模型的一个优点就是**我们可以借助它明确概率分布的参数的变化规律**。
- **专注于建模过程同样能实现对复杂现象的分析**。
- 统计模型让数据分析有了很大的进步，**堪称现代数据分析的标准框架**。

4.2 建模方法

- **响应变量（因变量）**是根据某个因素而变化（响应）的变量。
- **解释变量（自变量）**是对关注的对象的变化进行解释的变量。
- **参数模型**是尽量简化现象、用极少数参数表达的模型。在确定参数模型时，只确定少数参数即可。
- **非参数模型**不追求用尽量少的参数表达模型。非参数模型易于表达，但容易变得复杂，因而有时难以进行估计和解读。
- 在**线性模型**中，响应变量和解释变量之间的关系为线性关系。
- 统计模型中使用的参数叫作**系数**。在机器学习领域中，统计模型的系数也叫作**权重**。
- 建模分为两个步骤：**模型选择**和**参数估计**。
- **变量选择**就是为模型选取解释变量，没有解释变量的模型叫作**空模型**。
- 选择变量的两种方法：通过假设检验选择变量、通过**信息量**准则选择变量。
- 信息量准则可以量化所选模型与数据的契合度。**赤池信息量准则（**AIC**）**是一种常用的方法，

AIC 越小，模型越合适。

- 我们应当**在编写程序之前确定分析目的，并依据这个目的收集数据和建模**。

4.3 数据表示与模型名称

- 假设响应变量服从正态分布的线性模型叫作**正态线性模型**。正态线性模型属于参数模型。
- 在正态线性模型中，解释变量为连续变量的模型叫作**回归分析（回归模型）**。
- 含有多个解释变量的回归分析叫作**多元回归分析**，只有一个解释变量的回归分析叫作**一元回归分析**。
- 在正态线性模型中，解释变量为分类变量的模型叫作**方差分析模型**。
- 当解释变量为一个种类时叫作**一元方差分析**，当解释变量为两个种类时叫作**二元方差分析**。
- 响应变量未必服从正态分布的线性模型叫作**广义线性模型**。正态线性模型属于广义线性模型。
- 在机器学习领域中，响应变量为连续变量的模型叫作回归模型，正态线性模型属于广义的回归模型。响应变量为分类变量的模型叫作分类模型。

4.4 参数估计：最大似然估计

- 当参数为某值时抽到特定样本的概率（密度）叫作**似然（$\mathcal{L}$）**。
- 在给定参数时计算似然的函数叫作**似然函数**。
- 似然取对数就是**对数似然**。
- 求使得似然或对数似然最大的参数，并把这个参数作为参数估计值的方法就是**最大似然法**。
- 通过最大似然法估计得到的参数叫作**最大似然估计量（$\hat{\theta}$）**。
- 最大似然估计量对应的对数似然叫作**最大对数似然（$\ln \mathcal{L}(\hat{\theta})$）**。
- 与问题没有直接关系的参数叫作**多余参数**。
- 正态分布的参数有两个，分别为均值和方差。由于方差可以由均值求得，所以只要估计出均值，就可以间接得到方差。此时，方差按已知看待，无须再考虑。假设总体服从正态分布，在最大似然法中，方差σ^2就是多余参数。在对空模型进行估计时，只估计μ的值即可。

4.5 参数估计：最小化损失

- 在进行参数估计时，**损失函数**用于使参数最小。
- **残差**是响应变量的实际值与通过模型预测的值之间的差：$residual = y - \hat{y}$。
- 计算残差的平方并求和，得到的就是**残差平方和**：$\text{残差平方和} = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$。
- 求使得残差平方和最小的参数，并把这个参数作为参数估计值的方法就是**最小二乘法**。另一种说法是，最小二乘法是以残差平方和为损失指标，求使得损失最小的参数的方法。普通最小二乘法（OLS）。
- 最小二乘法得到的参数估计值等于假设总体服从正态分布时最大似然法的结果。
- 在机器学习领域中，**误差函数**就是对数似然的相反数。求最小的误差函数，就相当于求最大似然。于是最小二乘法还可以解释为，假设总体服从正态分布时让误差函数最小 。

4.6 预测精度的评估与变量选择

- **拟合精度**是模型与已知数据的契合度。
- **预测精度**是模型与未知数据的契合度。
- 拟合精度很高，预测精度却很低的现象叫作**过拟合**。模型过于契合已知数据是过拟合的原因。
- 解释变量过多是过拟合的常见原因。**删除多余的解释变量有可能提高预测精度**，而增加多余的解释

变量会提高拟合精度。

- 预测值和未知数据之间的误差叫作**泛化误差**。
- 用来估计参数的数据叫作**训练集**。
- 在估计参数时特意保留的一部分已知数据叫作**测试集**。
- 基于特定的准则把数据分为训练集和测试集，针对测试集评价模型预测精度的方法叫作**交叉验证法**。
- 交叉验证法主要有**留出交叉验证**和K**折交叉验证**两种。留出交叉验证从已知数据中取出p个数据作为测试集。K交叉验证把已知数据分为K组，取其中1组作为测试集，重复K次，以预测精度的均值作为最终的评估值。
- **赤池信息量准则（**AIC**）**的数学式为：$AIC=-2\times(\text{最大对数似然}-\text{参与估计的参数个数})$。$AIC$越小，模型越合适。$AIC$可以判断**增加的对数似然能否弥补更多的解释变量带来的缺点**。比起交叉验证法，AIC的一大优势是计算量更小。
- 统计模型的预测结果是一种概率分布，把它和数据真正服从的分布相比较，两者之间的差异用指标**相对熵**衡量：相对熵 $=\int g(x)\ln\frac{g(x)}{f(x)}dx=\int g(x)[\ln g(x)-\ln f(x)]dx$，其中，$g(x)$和$f(x)$为概率密度函数。将两个概率密度函数的对数差"$\ln g(x)-\ln f(x)$"看作期望值，可以更好地理解"相对熵是衡量概率分布的差异的指标"这句话的含义。
- 我们希望真实分布和所预测的分布之间的差距更小。$\int g(y)[\ln g(y)-\ln f(y)]dy\to\int[g(y)\ln g(y)-g(y)\ln f(y)]dy\to\int[-g(y)\ln g(y)]dy$，其相反数就是**平均对数似然**。
- 平均对数似然很难直接计算，因此我们经常使用最大对数似然代替它，这样就带来一个问题:最大对数似然有时远大于平均对数似然，所以二者有时会偏离过大。数学上已经证明，这个偏离的大小就是参与估计的参数的个数。因此，去掉这个偏离的结果就是AIC，这就是从最大对数似然中减去参与估计的参数个数的原因。
- 在假设检验没有得到希望的结果时换用AIC，这种行为与p值操纵是类似的。信息量准则中除了AIC，还有BIC、$AICc$等指标，我们同样不可以为了得到想要的结果而切换指标。

5 正态线性模型

5.1 含有单个连续型解释变量的模型（一元回归）

- AIC的核心是各个AIC之间的对比，其绝对值并不重要。通过相同做法计算出来的AIC的大小关系是不变的，只要不更换做法，就不会影响模型选择，这就意味着我们要避免跨工具计算AIC。
- 模型预测的响应变量的图形就是**回归直线**。当响应变量为连续变量时，它的图形叫作回归，这也是回归直线的名称来源。
- 非线性模型预测 的响应变量的图形叫作**回归曲线**。
- **决定系数**用来评估模型与已知数据的契合度，计算式为：$R^2=\frac{\sum_{i=1}^N(\hat{y}-\mu)^2}{\sum_{i=1}^N(y-\mu)^2}$，其中，$y$是响应变量，$\hat{y}$是模型的预测值，$\mu$是$y$的均值。如果响应变量的预测值和真实值相等，$R^2$就为1。
- 残差的变形可得：$y=\hat{y}+residual$，决定系数的分母可以分解为：$\sum_{i=1}^N(y-\mu)^2=\sum_{i=1}^N(\hat{y}-\mu)^2+\sum_{i=1}^N residual^2$，响应变量的差异等于模型可以预测的差异加上模型不可预测的残差平方和。既然存在这个关系，那么可以得到：$R^2=1-\frac{\sum_{i=1}^N residual^2}{\sum_{i=1}^N(y-\mu)^2}$。
- **修正决定系数**考虑了解释变量过多的惩罚指标，通过自由度修正了决定系数。解释变量越多，决定系数越大。决定系数过大会导致过拟合，因此需要对决定系数进行修正。修正决定系数的数学式

为：$R^2=1-\frac{\sum_{i=1}^{N}residual^2/(N-s-1)}{\sum_{i=1}^{N}(u-\mu)^2/(N-1)}$，其中，$s$为解释变量的个数。

- **分位图（Q-Q图）**是用来比较理论分位数与实际分位数的散点图。如果散点落在线上就表示数据服从正态分布。
- 要判断残差是否服从正态分布，还要观察**偏度（Skew）**和**峰度（Kurtosis）**的值。
- 偏度表示直方图左右非对称性的方向和程度。偏度大于0，则图形的右侧更宽。正态分布左右对称，所以它的偏度为0。偏度的数学式为：$Skew=E[\frac{(x-\mu)^3}{\sigma^3}]$。其中，$E()$为求期望值的函数，$x$为随机变量（此处为残差），$\mu$为$x$的均值，$\sigma$为$x$的样本标准差。
- 峰度表示直方图中心附近的尖锐程度。峰度越高，图形显得越尖锐。正态分布的峰度为3。峰度的数学式为：$Kurtosis=E[\frac{(x-\mu)^4}{\sigma^4}]$。
- 如果残差自相关，系数的t检验结果便不可信，这个现象叫作**伪回归**。

5.2 方差分析

- **方差分析（ANOVA）**是用来检验均值差的方法。
- 某些情况下，单纯使用t检验是行不通的。如果解释变量的水平大于2个，要检验各水平的均值之间是否存在显著性差异，就要使用方差分析。像天气状况、鱼的种类等分类变量就叫作**水平**。
- 要使用方差分析，数据的总体必须服从正态分布。另外，各个水平内部的方差必须相等。
- 反复检验导致显著性结果更易出现的问题叫作**多重假设检验**问题。
- 方差分析将数据的变化分为误差和效应，并据此计算统计量F**比**：F比 $=\frac{效应的方差}{误差的方差}$。如果F比的值大，就认为效应比误差的影响大。当总体服从同方差正态分布时，F比的样本分布就叫作F**分布**。
- 小提琴之间的高度差，即效应的大小，叫作**组间差异**。组间差异的自由度是水平数量减去1。
- 各个小提琴的高度，即误差的大小，叫作**组内差异**。组内差异的自由度是样本容量减去水平数量。
- 为了在建模时使用分类变量，我们引入了**虚拟变量**。
- 当解释变量为连续变量时，方差分析依然有效。在求F比之前，要定义自由度。当解释变量为连续变量时，组间差异的叫法变为**模型自由度**，组内差异的叫法变为**残差自由度**。
- 模型自由度为参与估计的参数个数减去1，残差自由度为样本容量减去参与估计的参数个数。
- 系数的t检验结果与方差分析的结果在解释变量只有1个时相等，在解释变量多于1个时不相等。

5.3 含有多个解释变量的模型

- 在$Type\ I\ ANOVA$中，如果改变解释变量的顺序，检验结果也会不一样。
- 在回归系数的t检验中，解释变量的顺序不会引起什么问题，然而多于2个水平就会出现多重假设检验问题。
- $Type\ II\ ANOVA$是方差分析的一种，它的结果不会因解释变量顺序的不同而不同。
- 在方差分析中，解释变量的效应是基于残差量化的。变量个数增加时所减少的残差平方和决定了变量的效应。这样一来，添加解释变量时的顺序就尤为重要了。
- $Type\ II\ ANOVA$根据解释变量减少时所增加的残差平方和量化解释变量的效应。即使解释变量的顺序不同，这种方法的结果也不会改变。通过这种方法得到的组间偏差平方和就叫作**调整平方和**。
- 如果使用AIC进行变量选择，就没必要像方差分析那样更换计算方法，直接建模并计算AIC即可。使用AIC进行变量选择的过程是比较固定的。它和系数的t检验不同，多水平的变量不会导致多重假设检验问题，所得模型的含义永远是“对未知数据的预测误差最小的变量组合”。AIC也没有检验的非对称性问题。
- 在解释变量之间相关性很强时出现的问题就是**多重共线性**。多重共线性问题最简单的解决方案就是去掉强相关变量中的一个。

6 广义线性模型

6.1 各种概率分布

- **二值随机变量**是只有两个值的随机变量。
- **伯努利试验**是得到两种结果中的一种的试验。
- 为了方便，把得到两种结果中的一种的概率称为**成功概率**。
- 在完成一次伯努利试验时得到的二值随机变量所服从的概率分布就是**伯努利分布**。
- 设成功概率为p，进行N次独立的伯努利试验，成功的次数m所服从的离散型概率分布叫作**二项分布**。服从二项分布的随机变量m的期望值为Np，方差为$Np(1-p)$。
- 二项分布的概率质量函数的数学式为：$Bin(m|N,p)=C_N^m\cdot p^m\cdot(1-p)^{N-m}$，其中，成功概率为$p$，试验次数为$N$，成功次数为$m$。形式的种类数可以使用组合的计算公式：$C_N^m=\frac{N!}{(N-m)!\cdot m!}$。
- **泊松分布**是“1个、2个”或“1次、2次”这样的**计数型数据**所服从的离散型概率分布。计数型数据全是自然数，与正态分布的实数数据不同。
- 泊松分布的参数只有1个，即强度λ。服从泊松分布的随机变量的期望值和方差都是λ。
- 泊松分布的概率质量函数的数学式为：$Pois(x|\lambda)=\frac{e^{-\lambda}\lambda^x}{x!}$，其中，$x$为计数型数据等离散型随机变量，$\lambda$为泊松分布的强度。
- 泊松分布可以由二项分布推导得出。成功概率趋近于0，试验次数趋向无穷大的二项分布就是泊松分布。
- **负二项分布**与泊松分布类似，是计数型数据所服从的分布，但它的方差大于泊松分布。例如，存在群居现象的生物个体数量的方差远超泊松分布的范围，这种现象叫作**过度离散**。此时使用负二项分布可以很好地将此类数据模型化。
- **伽马分布**与正态分布不同，它是非负的连续型随机变量服从的概率分布，方差的值随着均值的不同而变化（异方差）。
- 总体分布除了正态分布之外还可以使用其他概率分布的线性模型就是广义线性模型。这里，正态分布之外的概率分布属于**指数分布族**。
- 指数分布族的数学式为：$f(x|\theta)=exp[a(x)b(\theta)+c(\theta)+d(x)]$，其中，$x$为随机变量，$\theta$为概率分布的参数。

6.2 广义线性模型基础

- **线性预测算子**是线性关系式表示的解释变量。
- **联系函数**用于将响应变量和线性预测算子关联在一起，可以应用于响应变量。

6.3 $logistic$回归

- 概率分布为二项分布、联系函数为$logit$函数的广义线性模型叫作$logit$**回归**。
- $logit$函数的数学式为：$f(x)=\ln\frac{x}{1-x}$。
- $logistic$**函数（逻辑函数）**是$logit$函数的反函数，$logistic$函数的数学式为：$g(y)=\frac{1}{1+exp(-y)}$。
- 成功概率与失败概率的比值叫作**优势**，它表示是否容易成功，其数学式为：优势$=\frac{p}{1-p}$，其中，p为成功概率。
- 优势的对数叫作对数优势，$logit$函数也可以看作将成功概率转换为对数优势的函数。

- 优势的比值叫作**优势比**。优势比的对数叫作**对数优势比**。

6.4 广义线性模型的评估

- 二项分布的**皮尔逊残差**的计算式为：$Pearson\ residual = \frac{y-N\hat{p}}{\sqrt{N\hat{p}(1-\hat{p})}}$，其中，$y$为响应变量（二值随机变量），$N$为试验次数，$\hat{p}$为估计的成功概率。
- 皮尔逊残差的平方和叫作**皮尔逊卡方统计量**，是模型契合度的指标。
- **模型偏差**是评估模型契合度的指标。模型偏差越大，契合度越差。
- 二项分布的**偏差残差**的平方和就是模型偏差。
- 在很多机器学习的语境中，求$logistic$回归就是求使得**交叉熵误差**最小的参数。

6.5 泊松回归

- 概率分布为泊松分布、联系函数为对数函数的广义线性模型叫作**泊松回归**。

7 统计学与机器学习

7.1 机器学习基础

- **机器学习**是以让计算机拥有学习能力为目的的研究领域。
- 机器学习主要分为**有监督学习**和**无监督学习**。有监督学习研究的问题存在正确答案，无监督学习研究的问题不存在正确答案。
- **强化学习**解决的问题是在给定条件下寻找回报最大的行为。与有监督学习不同的是，强化学习研究的问题不存在正确答案。
- 按人们给定的规则输出预测结果的方法叫作**规则学习**，它和机器学习不是同一个概念。
- **统计模型的目的是理解获得数据的过程，机器学习的目的是通过计算得到未知数据。**

7.2 正则化、$Ridge$回归与$Lasso$回归

- 在参数估计中，向损失函数引入惩罚指标以防止系数过大的措施叫作**正则化**，惩罚指标叫作正则化项。在统计学中也将其叫作参数的**收缩估计**。
- $Ridge$**回归**将系数的平方和作为正则化项，这类正则化也叫$L2$**正则化**。
- $Lasso$**回归**将系数的绝对值之和作为正则化项，这类正则化也叫$L1$**正则化**。
- 在进行$Ridge$回归或$Lasso$回归之前，应当将解释变量标准化，即让解释变量的均值为0，标准差为1。

7.4 线性模型与神经网络

- 统计模型与机器学习中表示同一概念的术语可能不同。解释变量叫作**输入向量**；响应变量叫作**目标向量**；系数叫作**权重**；截距是值恒为1的解释变量，叫作**偏置**。
- **激活函数**用于将输入向量的加权和转换为输出。
- 输入向量进入的地方叫作**输入层**；输出预测值的地方叫作**输出层**；输入层与输出层中间的部分叫作**隐藏层（中间层）**。
- 由**多层感知机**组成的模型叫作**前馈神经网络（神经网络）**。含有多个隐藏层的模型也叫**深度学习**。