第一章　数据分析数学基础知识

党的二十大报告中有很多表述成果的数据，如国内生产总值从五十四万亿元增长到一百一十四万亿元，我国经济总量占世界经济的比重达百分之十八点五，提高七点二个百分点，稳居世界第二位；人均国内生产总值从三万九千八百元增加到八万一千元。谷物总产量稳居世界首位。十四亿多人的粮食安全、能源安全得到有效保障。城镇化率提高十一点六个百分点，达到百分之六十四点七。制造业规模、外汇储备稳居世界第一。建成世界最大的高速铁路网、高速公路网，机场港口、水利、能源、信息等基础设施建设取得重大成就。我们加快推进科技自立自强，全社会研发经费支出从一万亿元增加到二万八千亿元，居世界第二位，研发人员总量居世界首位。人均预期寿命增长到七十八点二岁。居民人均可支配收入从一万六千五百元增加到三万五千一百元。城镇新增就业年均一千三百万人以上。建成世界上规模最大的教育体系、社会保障体系、医疗卫生体系，教育普及水平实现历史性跨越，基本养老保险覆盖十亿四千万人，基本医疗保险参保率稳定在百分之九十五。及时调整生育政策。改造棚户区住房四千二百多万套，改造农村危房二千四百多万户，城乡居民住房条件明显改善。互联网上网人数达十亿三千万人。这些统计数据背后涉及诸多数理知识。通过本章大家可以了解关于数据分析的主要概率论与数理统计的基本原理；了解必然事件和随机事件的基础上理解频率和概率，并运用相关原理解决统计与数据分析相关问题。

第一节　随机事件及其概率

在我们日常生活中普遍存在的现象，主要分两类，一类是必然现象，另一类是或然现象。必然现象就是发生概率 100%，如磁石异性相吸；或然现象发生概率不是 100%，如股市涨跌、抛硬币。这些或然现象由于结果不能确定的现象叫随机现象。随机现象通过大量重复的随机试验进而使结果呈现一定的规律，这种规律就是概率论和数理统计的基础。

一、随机事件及概率

1. 随机事件

随机事件通常用大写字母表示，如 A、B、C，随机试验的每一个可能出现的结果称为基本事件或者叫样本点。所有的样本点（基本事件）的集合组成样本空间（一般用 Ω 表示）。由若干个基本事件组成的集合叫随机事件，随机事件是样本空间的一个子集。当随机事件集合中的某一基本事件发生，则代表这个随机事件的出现。

2. 事件的关系与运算

事件的关系主要包括并(和)、交(积)、差、互斥(互不相容)和逆(对立),运算分别如下。

(1) 事件的并(和)$A+B=A\cup B=${事件 A 和事件 B 至少出现一个}。

(2) 事件的并(和)$AB=A\cap B=${事件 A 和事件 B 都出现}。

(3) 事件的差 $A-B=${事件 A 出现但事件 B 不出现}。

(4) 事件的互斥 $AB=\varnothing$(空集符号){事件 A 和 B 不能同时出现}。

(5) 事件的逆(对立)$A+B=\Omega$,$AB=\varnothing$ {事件 A 和事件 B 必然且仅有一个出现}。

3. 随机事件的频率

随机事件的频率,通常用 F 或 f 表示。

为了衡量随机事件出现的可能性大小,可以用频率来评价。事件的频率是在相同的条件下,重复同样的试验 n 次,若事件出现了 m 次,则 m/n 是这 n 次试验中出现的频率。随机事件的频率不仅与试验次数 n 有关,而且还与试验的轮次有关。例如,在抛硬币中无法确定下一次正面出现的频率。

4. 随机事件的概率

随机事件的概率,通常用 P 或 p 表示。

我们为了更加精确地衡量随机事件发生可能性大小,需要用一个与试验次数和试验轮次无关的一个数字来表示,这个数字就是随机事件的概率。概率度量的是随机事件发生可能性的大小,进而可以反应随机现象的内在规律,那么概率究竟应该如何量化呢?

在多次重复同样试验的情况下,随机事件的频率呈现出稳定性,也就是在一个固定的数附近摆动,而这个固定的数就是概率。例如,历史上抛硬币试验的结果,费勒抛了 10 000 次,正面次数 4 979 次,频率 0.497 9;皮尔逊抛了 12 000 次,正面次数 6 019 次,频率 0.501 6;后来皮尔逊抛了 24 000 次,正面次数 12 012 次,频率 0.500 5。可见频率接近 0.5,而借助于极限的数学方法,可以确定这个固定的数为 0.5,也就是抛硬币出现正面的概率为 0.5。但是根据概率仍无法确定每一轮次随机事件发生与否,这也符合随机事件的内涵定义。

二、条件概率与事件的独立性

1. 条件概率

条件概率,通常用 $P(A|B)$表示。

日常在处理实际问题的时候,经常需要在已知部分事件发生的条件基础上来继续求解相关随机事件发生的概率,这样就需要引入条件概率。条件概率的概念是在某种条件下某件事发生的概率是多少。条件概率其实反映了一个朴素的概念,你平时在推理不确定的事情时,下意识地需要去找和事情有关的证据,因为证据越多,对不确定的事情的推理就越有信心,这些证据也就是条件。

设 A,B 是任意两个随机事件,且 $P(B)>0$(B 发生的概率大于 0),则在事件 B 发生的条件下,事件 A 发生的条件概率为 $P(A|B)=P(AB)/P(B)$。

例如,进行某一实验,所有的可能样本的集合为 Ω(所有样本),圆圈 A 代表事件 A 所

能囊括的所有样本，圆圈 B 代表事件 B 所能囊括的所有样本（图 1.1）。

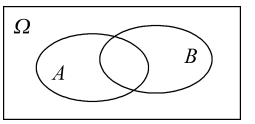

图　1.1

如图 1.1 所示，我们再来直观理解一下“事件 B 发生的条件下，事件 A 发生的条件概率”，“事件 B 发生的条件下”表明把样本的可选范围限制在了圆圈 B 中，也就是等价在圆圈 B 中 A 发生的概率，显然 $P(A|B)$ 就等于 AB 交集中样本的数目除以 B 的样本数目。那么为什么是样本的数目相除，而上面条件概率的公式是概率相除，其实等价于分子分母同除以总样本数，这就变成了条件概率公式。一般说到条件概率，事件 A 和事件 B 都是同一实验下的不同的结果集合，事件 A 和事件 B 一般是有交集的，若没有交集（AB 互斥），则条件概率为 0，因为 $P(AB)=0$，则 $P(A|B)=P(AB)/P(B)=0$。

2. 事件的独立性

随机事件 A 的发生和随机事件 B 的发生互不影响，就是事件 A 和 B 相互独立。对于任意随机事件 A 和 B，满足 $P(AB)=P(A)P(B)$，则称 A 和 B 相互独立。例如，两次抛硬币，并且相互独立也满足 $P(A|B)=P(A)$。

第二节　随机变量及其分布

为了进一步研究随机事件及其概率的规律，需要进行统一形式的定量数学处理，进而把随机事件的结果数量化处理，以便可以利用微积分相关理论和方法，由此引入随机变量和分布函数。

一、随机变量

1. 随机变量

随机变量通常用大写字母表示，如 X、Y、Z。

对于有些随机事件来说，其结果就是以数的形式出现的。例如，掷骰子出现的点数、考试的成绩、灯泡的寿命等。还有不以数形式的随机事件的结果，如抛硬币，但是我们可以用数字来定义结果。例如，正面朝上指定为 1，反面朝上指定为 0。这个过程就可以实现随机事件结果的数量化，而这种数量化的结果就是随机变量。这样我们就可以定义随机变量为在样本空间内对于每个样本点 a 总有一个实数 $X(a)$ 与之对应，那么可称这个实值函数 $X=X(a)$ 为随机变量。

随机变量按照取值是否有限可分为两种类型，如果随机变量的全部取值是可列的有限多个，那么为离散型随机变量；如果取值是不能列举，而是充满数轴的某一个区间，那么就是连续型随机变量。

2. 分布函数

分布函数通常表示为 $F(x)=P(X\leqslant x)$。

要想理解随机变量，除了要知道取值范围，更重要的还要了解其结果取值出现的分布规律，也就是概率分布。分布函数为刻画随机变量取值概率的函数，就是随机变量 X 落

入区间$(-\infty,x]$上的概率。通常用$F(x)=P(X\leqslant x)$来表示随机变量X的分布函数，记为$X\sim F(x)$，读作X服从$F(x)$。

对于离散型随机变量Y，若当Y取值y_1、y_2、y_3、…、y_n时候，其对应分布函数为$p_i=P(Y=y_i)$，$i=1,2,3,\cdots,n$，记为$X\sim\{p_i\}$。如常见的几何分布、二项分布(n重伯努利试验事件出现k次的概率)、泊松分布(单位时间内事件发生的次数)。

连续型随机变量结果的取值有无穷个实数，这些实数范围可以覆盖数轴上的某一区间或者整个数轴，无法通过列出取值概率来描述概率分布。我们引入概率密度函数来描述连续型随机变量的概率分布。这里我们利用微积分的知识对概率密度函数进行定义，设随机变量X的分布函数$F(x)$，若存在实数轴上一个非负可积函数$f(x)$，对于任意实数x，都有

$$F(x)=P(X\leqslant x)=\int_{-\infty}^{x}f(t)\mathrm{d}t$$

则称$f(x)$为X的概率密度函数。因为连续型随机变量X由其$f(x)$确定，所以可记为$X\sim f(x)$。

如图1.2所示，$P(a<x<b)=\int_{a}^{b}f(x)\mathrm{d}x$表示$X$落在区间$(a,b)$内的概率等于以阴影区域$(a,b)$和$f(x)$围成的面积。常见的连续型概率分布有均匀分布、指数分布、正态分布。

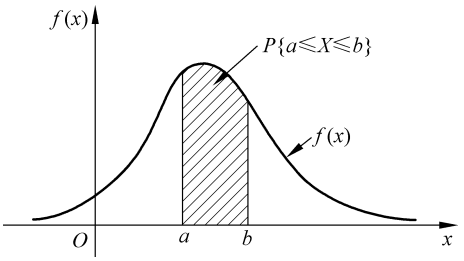

图 1.2

二、随机变量的数字特征

随机变量的概率分布能全面描述随机变量的统计特征，但是在一些时候我们只需要了解其中某些情况。例如，全班人的年龄、身高，现在需要知道平均年龄和平均身高，以及个体年龄与平均年龄的偏差程度，个体身高与平均身高的偏差程度，这些数字表示的随机变量的特征就是随机变量的数字特征。

1. 数学期望

数学期望，通常用大写字母$E(x)$表示。

数学期望表示随机变量所有可能取值的平均水平，严格意义上用随机变量以概率为权重的加权平均值来计算。随机变量的数学期望反映了随机变量的集中趋势，也就是在期望周围波动。

2. 二维随机变量函数的数学期望

若 X 和 Y 为 Ω 上的两个随机变量，则由 X 和 Y 构成的二维向量 (X,Y) 为二维随机变量。而二元函数 $F(x,y)=P(X\leqslant x,Y\leqslant y)$ 为联合分布函数。对于连续型随机变量，联合函数为 $F(x,y)=P(X\leqslant x,Y\leqslant y)=\int_{-\infty}^{x}\int_{-\infty}^{y}f(u,v)\mathrm{d}u\,\mathrm{d}v$。

设 $z=g(x,y)$ 是二元函数，

若为离散型 $(X=x_i,Y=y_j)\sim p_{ij}$，则数学期望可表示为

$$Eg(X,Y)=\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}g(x_i,x_j)p_{ij}$$

若为连续型 $(X,Y)\sim f(x,y)$，则数学期望可表示为

$$Eg(X,Y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}g(x,y)f(x,y)\mathrm{d}x\,\mathrm{d}y$$

3. 方差与标准差

方差与标准差，通常用 $\mathrm{Var}(x)$ 表示。

为了度量在期望周围波动的大小，也就是随机变量 X 的取值与 $E(X)$ 的偏离程度，引入方差和标准差的概念。

定义为设随机变量 X，若 $E[X-E(X)]^2$ 存在，则称 $\mathrm{Var}(X)=E[X-E(X)]^2$ 为随机变量的 X 的方差，$\sigma(X)=\sqrt{E[X-E(X)]^2}$ 为随机变量 X 的标准差。

4. 协方差

协方差，通常用 $\mathrm{Cov}(x)$ 表示。

协方差用于衡量两个变量的总体误差。而方差是协方差的一种特殊情况，即当两个变量是相同的情况。协方差表示的是两个变量的总体误差，这与只表示一个变量误差的方差不同。如果两个变量的变化趋势一致，也就是说如果其中一个大于自身的期望值，另外一个也大于自身的期望值，那么两个变量之间的协方差就是正值。如果两个变量的变化趋势相反，即其中一个大于自身的期望值，另外一个却小于自身的期望值，那么两个变量之间的协方差就是负值。期望值分别为 $E(X)$ 与 $E(Y)$ 的两个实随机变量 X 与 Y 之间的协方差 $\mathrm{Cov}(X,Y)$ 定义式为

$$\mathrm{Cov}(X,Y)=E[(X-E(X))(Y-E(Y))]$$

5. 相关系数

相关系数，通常用 r 或 ρ 来表示。

相关系数是最早由统计学家卡尔·皮尔逊设计的统计指标，是研究变量之间线性相关程度的量，一般用字母 r 表示。由于研究对象的不同，相关系数有多种定义方式，较为常用的是皮尔逊相关系数。用来度量两个变量间的线性关系。定义式为

$$r(X,Y)=\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

式中，$\mathrm{Cov}(X,Y)$ 为 X 与 Y 的协方差；$\mathrm{Var}(X)$ 为 X 的方差；$\mathrm{Var}(Y)$ 为 Y 的方差。相关系数定量地刻画了 X 和 Y 的相关程度，即 r 越大，相关程度越大；$r=0$ 对应相关程度

最低。

若 $0 \leqslant r \leqslant 1$ 称正相关，即两者变化方向一致。

若 $-1 \leqslant r \leqslant 0$ 称负相关，即两者变化方向相反。

6. 矩

在概率论与统计学中，矩是很重要的概念，以用来表征随机变量的数字特征，定义式为 $E\{[X-E(X)]^k\}$ 称为随机变量 X 的 k 阶中心矩，若 $E(X)=0$，则是 k 阶矩。一阶矩，指的是期望，表中心点的位置；二阶矩，指的是方差，表离散程度的度量（胖瘦）；三阶矩，指的是偏度，表分布偏斜方向和程度（歪斜）；四阶矩，指的是峰度，表在平均值处峰值高低（高矮）。

为了方便对比研究，我们经常做一个标准的线性变换，把中心点化为零（一阶矩归 0），离散程度化为 1（二阶矩归 1），这时在同样期望为 0 方差为 1 的标准情况下，随机变量的分布就可以用偏度和峰度来衡量了，也就是三阶矩和四阶矩。

第三节　中心极限定理与大数定律

在大量独立的重复试验中，有些随机变量可以表示为大量相互独立的和，为了刻画这些独立和的概率分布，下面将利用极限的思想来刻画这些分布。

一、中心极限定理

中心极限定理，是指概率论中讨论随机变量序列部分和分布渐近于正态分布的一类定理。这组定理是数理统计学和误差分析的理论基础，指出了大量随机变量近似服从正态分布的条件。它是概率论中最重要的一类定理，有广泛的实际应用背景。在自然界与生产中，一些现象受到许多相互独立的随机因素的影响，如果每个因素所产生的影响都很微小时，总的影响可以看作服从正态分布。中心极限定理就是从数学上证明了这一现象。最早的中心极限定理所研究的是在伯努利试验中，事件 A 出现的次数渐近于正态分布的问题。当样本量 N 逐渐趋于无穷大时，N 个抽样样本的均值频数逐渐趋近于正态分布，其对总体的分布不做任何要求，意味着无论总体是什么分布，其抽样样本的均值的频数分布都随着抽样数的增多而趋于正态分布。

二、大数定律

1. 切比雪夫不等式

19 世纪数学家切比雪夫研究统计规律中，论证并用标准差表达了一个不等式，这个不等式具有普遍的意义，被称作切比雪夫定理，其大意如下所述。

任意一个数据集中，位于其平均数 m 个标准差范围内的比例（或部分）总是至少为 $1-1/m^2$，其中 m 为大于 1 的任意正数。对于 $m=2$、$m=3$ 和 $m=5$ 有如下结果。

所有数据中，至少有 3/4（或 75%）的数据位于平均数 2 个标准差范围内。

所有数据中，至少有 8/9（或 88.9%）的数据位于平均数 3 个标准差范围内。

所有数据中，至少有 24/25(或 96%)的数据位于平均数 5 个标准差范围内。

切比雪夫不等式定义式为：设随机变量 X 存在数学期望 $E(X)$和方差 $D(X)$，则对于任意 $\varepsilon>0$，一直存在 $P[|X-E(X)|\geqslant\varepsilon]\leqslant\frac{D(X)}{\varepsilon^2}$。其实切比雪夫不等式想表达的是随机变量和期望的偏差符合统计规律，即偏差越大的范围的值占整体样本的概率就越小。其实就是整体的样本与期望的偏差不会太大。

2. 大数定律

概率论历史上第一个极限定理属于伯努利，后人称之为“大数定律”。概率论中讨论随机变量序列的算术平均值向随机变量各数学期望的算术平均值收敛的定律。根据这个定律知道，样本数量越多，则其均值就越趋近期望值。大数定律不会对已经发生的情况进行平衡，而是利用新的数据来削弱它的影响力，直至前面的结果从比例上看影响力非常小，可以忽略不计。这就是大数定律发生作用的原理。简而言之，大数定律发挥作用，是靠大数对小数的稀释作用。

在随机事件的大量重复出现中，往往呈现几乎必然的规律，这个规律就是大数定律。通俗地说，这个定理就是，在试验不变的条件下，重复试验多次，随机事件的频率近似于它的概率，偶然中包含着某种必然，是一种描述当试验次数很大时所呈现的概率性质的定律。

依概率收敛就是研究大数定律的基础，定义为设 X_n 是一个随机变量序列，c 是一个常数，对任意的 $\varepsilon>0$，有 $\lim\limits_{n\to\infty}P(|X_n-c|<\varepsilon)=1$，则称序列 X_n 依概率收敛于 c。若随机变量序列 X_n 的数学期望均存在，则对于任意 $\varepsilon>0$，有 $\lim\limits_{n\to\infty}P\left(\left|\frac{1}{n}\sum\limits_{i=1}^{n}X_i-\frac{1}{n}\sum\limits_{i=1}^{n}E(X_i)\right|<\varepsilon\right)=1$，则 X_n 服从大数定律。

大数定律的意义，首先就是以严格的数学形式表现了随机现象的一个性质，平稳结果(频率)的稳定性，还从理论上解决了，用频率近似代替概率和用样本均值近似代替理论均值的问题。

1) 切比雪夫大数定理

设 X_n 是一列相互独立的随机变量(或者两两不相关)，它们分别存在期望和方差($E(X)$和 $\mathrm{Var}(X)$)，方差满足一致有界，那么 X_n 服从大数定律。应用于抽样调查，就会有如下结论：随着样本容量 n 的增加，样本平均数将接近于总体平均数。从而为统计推断中依据样本平均数估计总体平均数提供了理论依据。特别需要注意的是，切比雪夫大数定理并未要求 X_n 同分布，相较于伯努利大数定律和辛钦大数定律更具一般性。

2) 伯努利大数定律

设 c 是 n 次独立试验中事件 A 发生的次数，且事件 A 在每次试验中发生的概率为 p，则对任意正数 ε，满足

$$\lim_{n\to\infty}P\left(\left|\frac{c}{n}-p\right|<\varepsilon\right)=1$$

该定律是切比雪夫大数定律的特例，其含义是，当 n 足够大时，事件 A 出现的频率将

几乎接近于其发生的概率，即频率的稳定性。在抽样调查中，用样本成数去估计总体成数，其理论依据即在于此。

3）辛钦大数定律

设 X_n 为独立同分布的随机变量序列，若 X_n 的数学期望 $E(X_k)$ 存在，则服从大数定律，即对任意的 $\varepsilon>0$，满足

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{k=1}^{n}X_k - E(X_k)\right|<\varepsilon\right)=1$$

总结一下，这两个定律都是在说样本均值性质。随着 n 增大，大数定律说样本均值几乎必然等于均值，指的是 n 只要越来越大，把这 n 个独立同分布的数加起来去除以 n 得到的这个样本均值会依概率收敛到期望值，但是样本均值的分布如何我们不知道。中心极限定律说，n 只要越来越大，这 n 个数的样本均值会趋近于正态分布，并且这个正态分布的方差越来越小。

第四节　统计量及抽样分布

在现实生活中，要研究的对象的概率分布往往是未知的，或者不完全知道。我们只能通过对所研究的对象或问题进行多次重复的试验与观察，已得到一定量的观测值，并且以这些数据为依据，对研究对象或问题进行推断。

一、总体、个体、样本和统计量

1. 总体、个体和样本

总体：所研究对象的全体为总体，可用随机变量 X 表示。

个体：组成总体的每一个元素称为个体，用随机变量 X_i 表示。

样本：从总体中随机抽取一部分个体称为样本，用多维随机变量（$X_1, X_2, X_3, \cdots, X_n$）来表示。

样本容量：样本中所含个体的数量。

样本观测值：一次具体样本抽取和测量得到的一组具体值。

简单样本：若总体 X 的样本满足相互独立、每一个 X_i 都与总体 X 同分布，则称为总体的简单随机样本。

2. 统计量

对于总体和样本的关系，是不能直接利用样本对总体进行推断，需要构造出与其相适应的样本函数，如果这个样本函数不含有任何未知参数，则称这个样本函数为统计量。如样本均值、样本方差、样本标准差、样本 k 阶原点矩、样本 k 阶中心矩。

二、常用统计量的抽样分布

由于统计量是随机变量，在使用样本统计量进行推断时，需要知道其统计量的概率分布，而统计量的分布称为抽样分布。

1. χ^2 分布(卡方分布)

χ^2 分布在数理统计中具有重要意义。χ^2 分布是由阿贝(Abbe)于 1863 年首先提出的,后来由海尔墨特(Hermert)和现代统计学的奠基人之一的卡尔·皮尔逊分别于 1875 年和 1900 年推导出来,是统计学中的一个非常有用的、著名的分布。

如果 X_n 相互独立来自总体 $N(0,1)$的样本,则记 $\chi^2=\sum_{i=1}^{n}X_i$ 为服从自由度为 n 的 χ^2 分布。密度函数的范围为$(0,+\infty)$,从图 1.3 可见当自由度 n 越大,密度曲线越趋于对称,n 越小,曲线越不对称。当 $n=1,2$ 时,曲线是单调下降趋于 0。当 $n\geqslant 3$ 时,曲线有单峰,从 0 开始先单调上升,在一定位置达到峰值,然后单调下降趋向于 0。

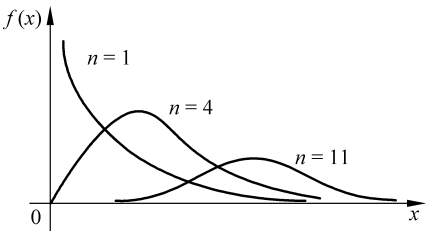

图　1.3

2. T 分布

T 分布是统计学家戈塞特(W. S. Gosset)在 1908 年以笔名 Student 发表的论文中提出的,故后人称为"学生氏(Student)分布"或"T 分布"。定义为设 X 和 Y 相互独立,且 $X\sim N(0,1)$,$Y\sim\chi^2(n)$,则 $T=\dfrac{X}{\sqrt{Y/n}}$服从自由度为 n 的 T 分布,如图 1.4 所示。用于根据小样本来估计呈正态分布且方差未知的总体的均值。如果总体方差已知(例如在样本数量足够多时),则应该用正态分布来估计总体均值。

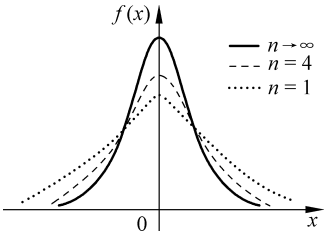

图　1.4

T 分布的密度函数与标准正态分布 $N(0,1)$密度很相似,它们都是关于原点对称,单峰偶函数,在 $x=0$ 处达到极大。但 t 分布的峰值低于 $N(0,1)$的峰值,T 分布的密度函数尾部都要比 $N(0,1)$的两侧尾部粗一些。

3. F 分布

F 分布是 1924 年统计学家罗纳德·费希尔(R. A. Fisher)提出,并以其姓氏的第一

个字母命名的。它是一种非对称分布，有两个自由度，且位置不可互换，如图 1.5 所示。F 分布有着广泛的应用。例如，在方差分析、回归方程的显著性检验中 F 分布都有着重要的地位。

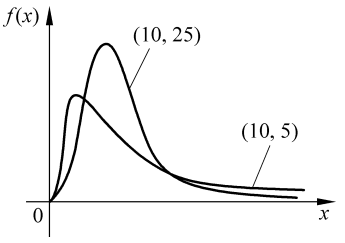

图 1.5

若总体 $N\sim N(0,1)$，X_m 与 Y_n 为来自 N 的两个独立样本，设统计量则称统计量 F 服从自由度 m 和 n 的 F 分布，记为 $F=\dfrac{X/m}{Y/n}$。

第五节　参数估计与假设检验

当我们所研究的对象或总体的分布已知，但一个或多个参数是未知的，这样我们可以借助总体的样本构造的统计量来估计未知参数，这就是参数估计，常包括点估计和区间估计。

另一种情况，在总体分布函数完全未知或不知其参数的情况下，为了推断总体的某些性质，可以提出某些关于总体的假设，再利用概率很小的事件在一次试验中可以认为基本上不会出现的原理进行验证假设的过程，就是假设检验。

一、参数估计

1. 点估计

设$(X_1,X_2,\cdots,X_n)$是来源于总体 X 的简单样本，θ 是总体 X 的一个未知参数，点估计就是找到一个适当的统计量 $\hat{\theta}(X_1,X_2,\cdots,X_n)$，用其试验的观测值 $\hat{\theta}(x_1,x_2,\cdots,x_n)$ 作为总体参数 θ 的近似值。点估计量是样本的函数，对于不同的样本值，点估计值一般不同。构造方法主要有矩估计法和最大似然估计法。

1）矩估计法

矩估计法就是利用样本的各阶原点矩来估计总体相同阶数的原点矩 $E(X^k)$，从而建立估计量满足的方程组，以解方程组求未知参数估计量的方法。

2）最大似然估计法

最大似然估计法最早是由高斯于 1821 年针对正态分布提出的一种参数估计方法，之后统计学家费希尔于 1922 年针对一般分布再次提出，使之成为一种普遍使用的构造点估计量的方法。最大似然估计法的思想就是如果在一次抽样中，样本的某个观测值出现，则

认为样本在这个观测值的取值的可能性最大，于是构建似然函数在该观测值处取最大值，进而可以作为未知参数的最大似然估计值。

设$(X_1,X_2,\cdots,X_n)$是来源于总体 X 的简单样本，观测值为 $x_1,x_2,\cdots,x_n$，若总体 X 为离散型随机变量，并且其分布概率可表示为 $P(X=x)=p(x,\theta)$，θ 为未知参数，则样本似然函数可表示为

$$L(\theta)=\prod_{i=1}^{n}p(x_i,\theta)$$

若总体 X 为连续型随机变量，并且其概率密度可表示为 $f(x,\theta)$，θ 为未知参数，则样本似然函数可表示为

$$L(\theta)=\prod_{i=1}^{n}f(x_i,\theta)$$

2. 区间估计

区间估计，是参数估计的一种形式。1934 年，由统计学家 J. 奈曼所创立的一种严格的区间估计理论。在点估计的基础上，给出总体参数估计的一个区间范围，该区间通常由样本统计量加减估计误差得到。与点估计不同，进行区间估计时，根据样本统计量的抽样分布可以对样本统计量与总体参数的接近程度给出一个概率度量。用数轴上的一段距离或一个数据区间，表示总体参数的可能范围。这一段距离或数据区间为区间估计的置信区间，是指在某一置信水平下，样本统计值与总体参数值间误差范围。按给定的概率值建立包含待估计参数的区间，其中这个给定的概率值称为置信度或置信水平，是总体参数值落在样本统计值某一区内的概率。置信区间越大，置信水平越高。划定置信区间的两个数值分别称为置信下限和置信上限，置信系数是这个理论中最为基本的概念。通过从总体中抽取的样本，根据一定的正确度与精确度的要求，构造出适当的区间，以作为总体的分布参数(或参数的函数)的真值所在范围的估计。

假设 θ 是总体 X 中的未知参数，从其样本找出两个统计量 θ_1 和 $\theta_2(\theta_1<\theta_2)$，使得对于任意的 $\alpha(0<\alpha<1)$，都满足 $P(\theta_1<\theta<\theta_2)=1-\alpha$，则称区间$(\theta_1,\theta_2)$为 θ 的置信区间，$1-\alpha$ 为置信水平，α 为显著性水平。

置信区间(θ_1,θ_2)是随机区间，其统计含义为：在相同条件下抽样 N 次，每次抽样的样本容量均为 n。由于每次得到的样本值不同，故每次得到的区间也可能不同，这样就得到了 N 个区间。对每个区间而言，它要么包含 θ 的真值要么不包含 θ 的真值，根据伯努利大数定律，在这 N 个区间中，包含 θ 真值的区间约占 $100(1-\alpha)\%$，不包含 θ 真值的区间约占 $100\alpha\%$。例如，若 $\alpha=0.01$，$N=1\,000$ 次，那么得到 1 000 个区间中不包含 θ 真值有 10 个。

通过图 1.6 正态分布图和图 1.7 卡方分布图可以更直观地看到置信区间和置信水平在坐标轴上的位置。

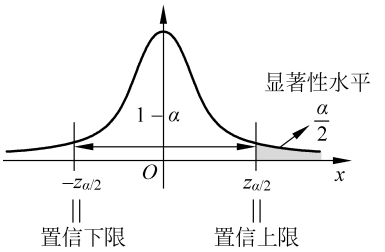

图 1.6

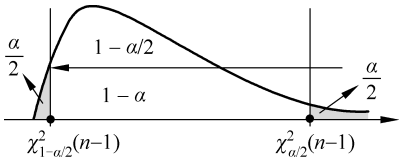

图 1.7

二、假设检验

1. 假设检验的基本原理和概念

假设检验的基本思想是小概率反证法思想。小概率思想是指小概率事件($P<0.01$ 或 $P<0.05$)在一次试验中基本上不会发生。反证法思想是先提出假设(检验假设 H_0),再用适当的统计方法确定假设成立的可能性大小,若可能性小,则认为假设不成立,若可能性大,则还不能认为假设不成立。也就是欲检验其正确性的为零假设,零假设通常由研究者决定,反映研究者对未知参数的看法。相对于零假设的其他有关参数之论述是备择假设,它通常反映了执行检定的研究者对参数可能数值的另一种(对立的)看法。如果我们检测的样本数据导致了一个小概率事件的发生,则拒绝原假设 H_0,如果没有发生小概率事件,则接受原假设 H_0,至于这个小概率的标准我们需要事先指定好,常用的一般为5%和1%,用 α 来表示,称为显著性水平。当检验统计量落入小概率区域(拒绝域),拒绝域的边界点为临界点。

有时,根据一定的理论或经验,认为某一假设 H_0 成立。例如,通常有理由认为特定的一群人的身高服从正态分布。当收集了一定数据后,可以评价实际数据与理论假设 H_0 之间的偏离,如果偏离达到了“显著”的程度就拒绝 H_0,这样的检验方法称为显著性检验。偏离达到显著的程度通常是指定一个很小的正数 α(如 0.05,0.01),即当 H_0 正确时,它被拒绝的概率不超过 α,称 α 为显著性水平。这种假设检验问题的特点是不考虑备择假设,考虑实验数据与理论之间拟合的程度如何,故此又被称为拟合优度检验。拟合优度检验是一类重要的显著性检验。

总而言之,假设检验是对总体分布中的某些参数做出假设,根据样本值提供的信息。

应用概率性质的反证法，检验这种假设是否成立，这一统计推断过程称为参数的假设检验。过程一般包括以下步骤。

(1) 为了检验一个零假设(即虚拟假设)是否成立，先假定它是成立的，然后接受这个假设之后，是否会导致不合理结果。如果结果是合理的，就接受它；如不合理，则否定原假设。

(2) 所谓导致不合理结果，就是看是否在一次观察中，出现小概率事件。通常把出现小概率事件的概率记为 α，即显著性水平。它在次数函数图形中是曲线两端或一端的面积。因此，从统计检验来说，就涉及双侧检验和单侧检验问题。在实践中采用何类检验是由实际问题的性质来决定的。一般可以考虑以下检验方法。

① 双侧检验。如果检验的目的是检验抽样的样本统计量与假设参数的差数是否过大，就把风险平分在右侧和左侧。比如显著性水平为 0.05，即概率曲线两侧各占一半，即 0.025。

② 单侧检验。这种检验只注意估计值是否偏高或偏低。如只注意偏低，则临界值在左侧，称左侧检验；如只注意偏高，则临界值在右侧，称右侧检验。

对总体的参数的检量，是通过由样本计算的统计量来实现的。所以检验统计量起着决策者的作用。

2. *P* 值法

P 值是用来判定假设检验结果的一个参数，由费希尔首先提出。*P* 值就是当原假设为真时所得到的样本观察结果或更极端结果出现的概率。如果 *P* 值很小，说明原假设情况的发生的概率很小，而如果出现了，根据小概率原理，我们就有理由拒绝原假设，*P* 值越小，我们拒绝原假设的理由越充分。总之，*P* 值越小，表明结果越显著。但是检验的结果究竟是“显著的”还是“高度显著的”需要根据 *P* 值的大小和实际问题来决定。

在假设检验中，首先认定原假设是对的，然后可以据此假设建立一个概率分布。在此概率分布下，以本次抽样实际情况为起点，一直到最极端情况所在区间的总概率，就是所谓的 *P* 值。如果原假设是对的，那么 *P* 值应该不会太小。如果 *P* 值太小，毕竟本次实际抽样的结果实际发生了，那么就是小概率实际居然发生了(小概率事件实际一次试验几乎不会发生的)，可以反证出原假设是不成立的(概率反证法)。实际中，所见的 *P* 值一般都很小，往往拿 0.05 作为界限，如果 *P* 值小于 0.05，表示基于当前抽样原假设不成立。

【思考】

1. 什么是频率？什么是概率？两者的区别是什么？
2. 什么是中心极限定理和大数定律？举个现实中的例子。
3. 简单描述一下假设检验的思想。

【扩展阅读】

什么是概率？

【即测即练】

自学自测  扫描此码