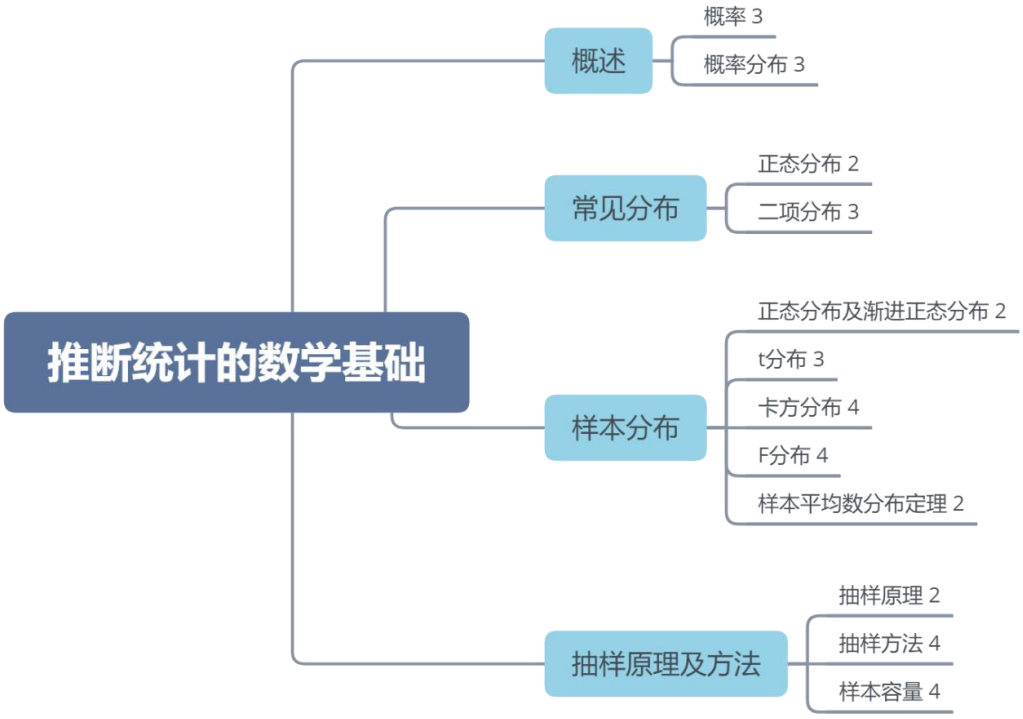
第六章　推断统计的数学基础



一、概述

（一）概述

概率是表明随机事件出现可能性大小的客观指标。

1.分类

（1）后验概率

对随机事件进行 n 次观察，某一事件 A 出现的次数 m 与观测次数 n 的比值。当 n 趋近无穷时，这个比值将稳定在一个常数 P 上，这一常数称作概率，P(A)= m/n

（2）先验概率

在满足试验可能结果数有限且每一种结果出现的可能性相等的条件下，随机事件包含的可能结果数（m）除以可能结果总数（n）。P（A）= m/n。

当进行多次观测时，按观测结果计算的概率（后验概率）基本接近先验概率。

2.性质

（1）公理系统

①任一随机事件 A 的概率均为非负；

②一定条件下必然事件的概率为 1；

③一定条件下不可能事件的概率为 0。

后两个公理的逆定理不成立。

（2）加法公理

P（A +B）=P（A）+ P（B）=P（A 或 B）

（3）乘法公理

P（AB）=P（A）X P（B）=P（A 且 B）

3.随机取样的要求

（1）在总体中的每个个体都有相等的机会被选择。

（2）返还取样。这对许多统计公式都是必要的，如果选择不只一个个体，选择这个个体的概率与选择另一个个体的概率必须相同。

（二）概率分布

对随机变量取值的概率分布情况用函数进行描述，依据不同标准可分为不同类型：

1.按变量描述的数据特点划分，分为离散分布和连续分布。

2.按函数的来源划分，分为经验分布和理论分布。

3.按描述数据的特征分，分为基本随机变量分布（原始数据的分布）和抽样分布（样本统计量的分布）。

二、常见分布

（一）正态分布

又称常态分布，由棣•莫弗发现，拉普拉斯、高斯也对其研究作了贡献，有时又称高斯分布。它是连续随机变量概率分布的一种，是应用最广泛的一种理论分布，记作 N（μ，σ2）。

1.特点

（1）呈对称分布（对称不一定是正态）。在正态分布中，均值、中数、众数相等。 正态曲线的形状像一口钟，两头小，中间大，大部分的原始分数都集中分布在均值附近，极端值相对而言比较少。

（2）中央点最高，曲线先向内弯后向外弯，拐点在土 1 个标准差处。两端向靠近横轴处不断延伸，但始终不会与横轴相交。

（3）正态分布是一族分布，即其形态会随随机变量的均值和标准差的变化而变化。所有正态分布都可以经由 Z 分数转换成标准正态分布，标准正态分布的μ=0，σ2=1，记作 N（0，1）正态分布，具有固定的形态。另外，标准正态分布因具有固定的均值和标准误，所以可以排除不同样本数据单位不同造成的混乱，更易进行推断分析。

（4）正态曲线下面积为 1，分布下包含了所有数据。

（5）正态分布曲线下，标准差和面积有一定数量关系。土 1 个标准差包含总面积的 68.26%，+1.96 个标准差包含总面积的 95%，+2.58 个标准差包含总面积的 99%。

（6）正态分布下各差异量数之间有固定比率。

2.正态分布的应用

（1）正态分布表的应用

①已知概率（P）可查 Z 分数。

②已知 Z 分数可查概率（P）。

③已知概率（P）或 Z 分数可查密度值（y）。

（2）正态分布在研究中的应用

①化等级评定为测量数据。

②确定测验题目的难易度。

③按能力分组，确定人数。

④测验分数的正态化。

（二）二项分布

试验仅有两种不同性质结果的概率分布，可以说是两个对立事件的概率分布。

1.二项试验

又称贝努里试验，需满足：

（1）任何一次试验恰有两个结果。

（2）共有 n 次试验，n 是预先给定的任一正整数。

（3）各次试验各自独立。

（4）某结果的概率在任一次试验中都是固定的。

2.特点

（1）二项分布是离散型分布

①p=q 时图形对称。

②p≠q 时成偏态，若 n 很大，则二项分布接近于正态分布。

（2）若二项分布满足 p<q，且 np≥5，或 p>q 且 nq≥5，二项分布接近正态分布，μ=np，$\sigma=\sqrt{npq}$，n 为独立试验的次数，p 为成功事件的概率，q=1-p。

3.应用：主要用于解决含有机遇性质的问题。

三、样本分布

样本分布即样本统计量的分布，只有知道了样本分布，才能依据样本对总体进行推论。

（一）正态分布

1.样本均值的分布

（1）总体为正态，方差已知，样本均值的分布为正态分布。

（2）总体非正态，方差已知，但是样本足够大（n>30），样本均值的分布为渐近正态分布。

这两种情况下的样本均值分布的均值和标准差：

$$\mu_{\bar{X}}=\mu,\ \delta_{\bar{X}}=SE=\frac{\delta}{\sqrt{n}}$$

$\mu_{\bar{X}}$ 为平均数的平均数；$\delta_{\bar{X}}$为平均数的标准差，又称标准误，它估计了由于随机性所造成的样本平均数与总体平均数之间的标准差量。

2.样本方差及标准差的分布

样本足够大时（n>30），样本方差及标准差渐趋于正态分布，这时样本方差或样本标准差分布的均值和标准差是：

$$\overline{X_S}=\delta,\ \overline{X_{S^2}}=\delta^2,\ \delta_s=\frac{\delta}{\sqrt{2n}}$$

（二）t 分布

1.样本均值分布

（1）总体为正态，方差未知，样本均值的分布为 t 分布。

（2）总体非正态，方差未知，但是样本足够大（n >30）， 样本均值的分布近似为 t 分布。这两种情况下：

$$\mu_{\bar{X}} = \mu,\ \delta_{\bar{X}} = SE = \frac{S}{\sqrt{n-1}},\ df = n-1$$

2.特点

t 分布是一种左右对称、峰态较高狭，形状随自由度 n-1 的变化而变化的一族分布。有以下特点:

（1）形状：以均值为 0 左右对称，左侧 t<0，右侧 t>0。

（2）取值：变量取值在[一∞，+∞]。

（3）变化

n→+∞时，t 分布为正态分布，方差为 1，其中：

①若 df=n-1>30 时，t 分布接近正态分布，方差大于 1，随 df 增大，方差渐趋于 1。

②若 df=n-1<30 时，t 分布与正态分布相差较大，随 df 减小，方差变大，分布中间变低、尾部变高。

（三）卡方分布

从正态总体中随机抽取无限多个数量为 n 的随机变量，这些变量的平方和或者标准分数的平方和的分布即为卡方分布。

卡方分布有以下特点：

1.形状

是一个正偏态的一族分布，属于连续型分布（有些离散型分布也近似卡方分布）。

2.取值

$$\chi^2 = \frac{\sum (X-\bar{X})^2}{\delta^2},\ df = n;\ \text{若}\mu\text{未知},\ \chi^2 = \frac{\sum (X-\bar{X})^2}{\delta^2},\ df = n-1$$

均为正值。

3.变化

n 或 n-1 越小，分布越偏斜；df→+∞时，为正态分布。

4.其他

（1）卡方分布的和也是卡方分布，即卡方分布具有可加性。

（2）df>2 时，μ_{χ^2} =df，$\delta^2_{\chi^2}$=2df；一般情况下，df 值均大于 2。

（四）F 分布

从两个正态分布总体中随机抽取容量为 n_1、n_2 两个样本，计算χ^2 值，每个χ^2 随机变量除以对应的自由度 df_1 与 df_2 之比，称为 F 比率，这无限多个 F 值的分布即为 F 分布。

$$F = \frac{\chi_1^2}{\chi_2^2} / \frac{df_1}{df_2}$$

若两样本取自同一总体，此时可将上式化简为（方差齐性检验）：

$$F = \frac{S^2_{n_1-1}}{S^2_{n_2-1}}$$

F 分布有以下特点：

1.形状

F 分布是一个正偏态的一族分布。

2.取值

F 值总为正值（方差之比）。

3.变化

F 分布随着分子分母的自由度的增加而渐趋于正态分布。

4.其他

分子自由度为 1 时，分母自由度为任意值，都有 F 值与分母自由度相等概率的 t 值（双侧概率）的平方相等。$F=t^2$ （两种处理水平）。

（五）样本平均数分布定理

1.中心极限定理

对于任意平均数为μ、标准差为σ的总体，样本容量为 n 的样本平均数分布的平均数为μ，标准差为σ/ $\sqrt{n}$。n>30 或趋于无穷大时，样本平均数的分布趋近于正态分布。

2. 大数定律

样本容量 n 越大，$\overline{X}$与μ接近的可能性越大，标准误越小，即样本越能代表总体。

四、抽样原理和抽样方法

（一）抽样原理

1. 随机性原则

（1）每个个体被选取的概率相等。

（2）进行返回取样，以保证每个个体每次被抽取的概率不变。

这样就有很大的可能性使得样本保持和总体相同的结构，即可以保证样本代表总体。

2.最大允许抽样误差 d

样本均值$\overline{X}$与总体均值μ之间的差异若超过最大允许抽样误差（d），则说明$\overline{X}$已经不是来自总体μ的一个样本了。可以通过改变样本容量 n 来控制 d，也可以根据 d 来推算 n。

$$d = Z \times SE_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

（二）抽样方法

1.简单随机取样法

包括抽签法、随机数字法，适合总体数目较小、个体差异较小时用。

最能体现随机化原则，但在大规模抽样时费时费力。

2.等距取样法

又称系统抽样、机械抽样，适合总体数目庞大时用。抽样方法简单，样本代表性较强；但当总体具有某一种周期性变化时，可能忽略已有信息。

3. 分层随机取样法

要求层内尽量同质，层间尽量异质，个体差异较大时用。

（1）按各层人数比例分配：人数多的层多分配，人数少的层少分配。

（2）最佳分配：标准差大的层多分配，标准差小的层少分配。

4.多段随机取样法

如两阶段随机取样，总体容量很大时用。

（三）样本容量

1.样本容量与总体没有固定的数量关系；

2. 一般来说，样本容量越大，产生的误差越小；

3.若总体差异很大，则需要增大样本容量以减少抽样误差；

4.在已知置信区间和置信水平的前提下，可以通过公式和查表的方式计算或找到对应的样本量大小。