概率论与数理统计知识点整理

统计推断部分

KujoStar-hmc

2022 年 1 月 11 日

○、统计中的相关概念

1. 统计总体

统计所研究的总体可以用一个概率分布来表示。总体分为无限总体、有限总体和虚拟总体。当个体数量很多时，可以近似看作无限总体进行研究。

2. 统计模型

在研究总体的过程中，我们一般通过一族概率分布建立数学模型，例如常见的正态分布族或者指数分布族。有时候，我们有有限个参数就可以成功建立概率分布族，例如一系列正态分布 $N(\mu_i, \sigma_i^2)$；但是更多的时候，我们仅仅能保证建立连续型随机变量的数学模型，使其期望存在、方差有限，而不能用有限个参数来刻画。这就是非参模型，进一步会涉及到模型近似的问题。

3. 样本

样本是一组随机变量 $(X_1, X_2, \cdots, X_n)$，其中每个 X_i 都来自总体，n 称为样本容量。

简单随机抽样

我们定义简单随机抽样：总体个数 N 有限，从总体中无放回抽取得到样本，任意容量相等的样本都有相同的发生概率，设容量为 k，则该容量的样本的发生概率为

$$p = \frac{1}{\binom{N}{n}}$$

随机样本

设 $X_i (i = 1, 2, \cdots, n)$ 为相互独立且同分布的一组随机变量，可以看作来自同一总体的容量为 n 的一个样本。可以看作是从样本中进行有放回的抽取得到。

4. 样本统计量

设一个样本为 $(X_1, X_2, \cdots, X_n)$，则统计量是样本的一个函数 $T(X_1, X_2, \cdots, X_n)$，完全由样本的情况决定，是一种用于简化数据方便研究的方式。

常见的样本统计量

1. 样本均值:
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
2. 样本方差:
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$
3. 设 μ 为总体均值，考虑变量 $\overline{X} - \mu$。若 μ 未知，则其值取决于总体，此时 $\overline{X} - \mu$ 不是统计量。若 μ 已知，则 $\overline{X} - \mu$ 是统计量。

5. 统计推断

统计推断的过程是一个总体与样本相互作用的过程。根据不同的抽样方法，从总体中产生特定的样本分布；从已有的样本分布，经过估计与检验等方法，可以得到总体分布的某些性质。这样的研究就是统计推断。

一、参数的点估计

一般情况下，我们研究的总体的概率分布中会有若干个未知的参数。通过一定的方法对已知的样本进行处理，得到这些未知参数的估计值，叫做参数的点估计。

1. 矩估计

定义

我们首先定义样本矩。设 $X_1, X_2, \cdots, X_n$ 为来自同一总体的样本，彼此独立同分布，则可以定义样本的 k 阶原点矩 a_k 和 k 阶中心矩 m_k:

$$a_k := \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

$$m_k := \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k$$

利用大数定律（LLN），我们可以如下近似:

$$a_k \longrightarrow E(X^k)$$

$$m_k \longrightarrow E((X-\mu)^k)$$

这里的 X 代表一个服从总体概率分布的随机变量。这样，我们就可以利用样本矩进行参数的矩估计。

典例

1. 设总体分布为 $N(\mu,\sigma^2)$，μ 与 σ 均未知。已知样本容量为 n 的样本 $X_i(i=1,2,\cdots,n)$，则可以进行如下的矩估计：

$$\mu = E(X) \approx a_1 = \overline{X}$$

$$\sigma^2 = Var(X) \approx m_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

从中我们可以看出，总体分布的期望可以用一阶样本原点矩 a_1 估计，总体分布的方差可以用二阶样本中心矩 m_2 估计。总体分布中的参数绝大部分情况下都与期望与方差这些数字特征有关，从而我们可以列出方程组得到参数的估计。

2. 设总体分布为 $Exp(\lambda)$，λ 未知。已知样本容量为 n 的样本 $X_i(i=1,2,\cdots,n)$，则可以进行如下的矩估计：

$$\frac{1}{\lambda} = E(X) \approx a_1 \Longrightarrow \lambda = \frac{1}{a_1}$$

$$\frac{1}{\lambda^2} = Var(X) \approx m_2 \Longrightarrow \lambda = \sqrt{\frac{1}{m_2}}$$

可以看到这个例子中对参数 λ 有两种矩估计方式。一般地，我们尽量采用低阶矩进行参数估计，这样在实际应用中具有更好的计算稳定性。

2. 极大似然估计（MLE）

定义

对于观测 $(X_1,X_2,\cdots,X_n)$，我们假设具有联合分布 $f(x_1,x_2,\cdots,x_n;\theta)$，其中 θ 为待估计的参数。我们定义观测的似然函数为：

$$L := f(X_1,X_2,\cdots,X_n;\theta)$$

由于这里的 X_i 都是已知的观测量，则 L 为关于参数 θ 的函数，可以记作

$$L = L(\theta)$$

我们定义

$$\theta^* := \underset{\theta}{\operatorname{argmax}}\, L(\theta)$$

也即 θ^* 是函数 $L(\theta)$ 取得最大值时自变量的值。我们称 θ^* 为参数 θ 的极大似然估计。

相关补充

1. 若 X_i 之间相互独立同分布，设总体的分布为 $f_1(x;\theta)$，则我们有

$$L(\theta) = f(X_1,X_2,\cdots,X_n;\theta) = f_1(X_1;\theta)f_1(X_2;\theta)\cdots f_1(X_n;\theta)$$

2. 由定义我们可以知道，θ^* 的取值是依赖于已有的观测值 X_i 的。

3. 一般来说，为了便于计算，我们通常会先对似然函数 L 取自然对数，得到 $L' = \ln L$，之后研究函数 L' 的最大值。原因是取自然对数后不改变原有的最大值点，同时可以将似然函数中可能存在的乘积（例如独立同分布时的似然函数）转化为相加，便于求导得出最大值点。

4. 极大似然估计具有不变性，也即对任意实函数 g，都有 $[g(\theta)]^* = g(\theta^*)$。

5. 极大似然估计的结果与矩估计的结果没有必然联系。

典例

1. 设 $X_i (i = 1, 2, \cdots, n)$ 为来自总体 $N(\mu, \sigma^2)$ 的独立同分布样本，μ, σ 未知。则根据定义，我们有似然函数：

$$L = L(\mu, \sigma) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X_i - \mu)^2}{2\sigma^2}} \right)$$

根据极大似然估计方法，对 $L(\mu, \sigma)$ 求自然对数，之后分别对 μ, σ^2 求偏导，得到极大似然方程：

$$\frac{\partial \ln L}{\partial \mu} = \frac{\partial \ln L}{\partial (\sigma^2)} = 0$$

从而可以得到极大似然估计：

$$\mu^* = \overline{X}, (\sigma^2)^* = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

由 MLE 的不变性，有

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

可以利用 Hesse 矩阵等方法验证此时函数 L 确实取得最大值。

2. 设 $X_i (i = 1, 2, \cdots, n)$ 为来自总体 $U(0, \theta)$ 的独立同分布样本，θ 未知。则根据定义，我们有似然函数：

$$L = L(\theta) = \begin{cases} \frac{1}{\theta^n}, & \forall i, X_i \in [0, \theta] \\ 0, & \exists i, X_i \notin (0, \theta) \end{cases}$$

我们发现无法通过取对数求导的方式求得极大值点。从定义入手，我们发现当 θ 取到取值范围内的最小值时，L 可以取到最大值。则我们显然有 $\theta \geq \max\{X_1, X_2, \cdots, X_n\}$，于是立得此时的极大似然估计为

$$\theta^* = \max\{X_1, X_2, \cdots, X_n\}$$

3. 矩估计与 MLE 相关评述

1. 参数估计的方法不是唯一的，无法判定绝对的优劣。

2. 极大似然方程的根并不一定是所求的极大似然估计值，需要验证是否是极大值点（虽然实际题目中基本都是确定的）。

3. MLE 需要知道总体的 pdf/pmf，而矩估计只需要知道参数和矩的关系。

二、点估计的性质

1. 无偏性

定义

我们首先定义一个估计 $\hat{\theta}$ 的偏差（Bias）：

$$E_\theta(\hat{\theta}-\theta)=E_\theta(\hat{\theta})-\theta$$

这里 θ 当作一个确定的常数。若估计的偏差为 0，也即对 $\forall\theta$，都成立

$$E_\theta(\hat{\theta})=\theta$$

则称 $\hat{\theta}$ 为 θ 的一个无偏估计。一般地，称 $\hat{g}(X_1,X_2,\cdots,X_n)$ 是 $g(\theta)$ 的一个无偏估计，等价于成立

$$E\left(\hat{g}(X_1,X_2,\cdots,X_n)\right)=g(\theta)$$

典例

1. 设总体分布的期望为 μ，方差为 σ^2，设 $X_i(i=1,2,\cdots,n)$ 为来自总体的独立同分布样本。若用样本均值（一阶原点矩）$\overline{X}$ 估计 μ，由于 $E(\overline{X})=\mu$，则这一估计是无偏估计；若用二阶中心矩 m_2 估计 σ^2，计算可得

$$E(m_2)=E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\right)=\frac{n-1}{n}\sigma^2$$

则这一估计为有偏估计，系统性偏小。同时易得样本方差 $S^2=\dfrac{1}{n-1}\sum\limits_{i=1}^{n}(X_i-\overline{X})^2$ 是 σ^2 的一个无偏估计。

2. 设总体的分布为 $U(0,\theta)$，$X_i(i=1,2,\cdots,n)$ 为来自总体的独立同分布样本。我们知道 θ 的极大似然估计为 $\theta^*=\max\{X_1,X_2,\cdots,X_n\}$，经过计算可得

$$E(\theta^*)=\frac{n}{n+1}\theta$$

则 θ^* 为系统性偏小的无偏估计。计算过程可以参考作者的知乎文章：前往奇妙世界，将 θ^* 的 pdf 求出来再积分求期望即可。

2. 均方误差准则（有效性准则）

定义

我们首先定义一个估计 $\hat{\theta}$ 的均方误差：

$$E_\theta\left((\hat{\theta}-\theta)^2\right)=Var(\hat{\theta}-\theta)+E_\theta^2(\hat{\theta}-\theta)=Var(\hat{\theta})+E_\theta^2(\hat{\theta}-\theta)$$

对于无偏估计 $\hat{\theta}$，其均方误差即为 $Var(\hat{\theta})$。由此，我们有均方误差准则：

设 $\hat{\theta}_1, \hat{\theta}_2$ 均为 θ 的无偏估计，若对 $\forall\theta$ 均成立

$$Var(\hat{\theta}_1) \leq Var(\hat{\theta}_2)$$

且 $\exists\theta_0$，使得

$$Var_{\theta_0}(\hat{\theta}_1) < Var_{\theta_0}(\hat{\theta}_2)$$

则称估计 $\hat{\theta}_1$ 再均方误差意义下优于 $\hat{\theta}_2$。

典例

我们需要估计总体均值 μ，现有独立同分布的样本 $X_i(i = 1, 2, \cdots, n)$，易得 $\overline{X}$，X_1，$\dfrac{X_1 + X_2}{2}$ 均为 μ 的无偏估计。

设总体的方差为 σ^2，我们经过计算易得

$$Var(\overline{X}) = \frac{\sigma^2}{n}$$

$$Var(X_1) = \sigma^2$$

$$Var\left(\frac{X_1 + X_2}{2}\right) = \frac{\sigma^2}{2}$$

由均方误差准则，可得上述三个估计中均方意义下最优的估计是 $\overline{X}$。

推广

设 $\hat{\theta}_0$ 是参数 θ 的一个无偏估计，且对其他任意的无偏估计 $\hat{\theta}$ 均成立

$$Var(\hat{\theta}_0) \leq Var(\hat{\theta})$$

则称 $\hat{\theta}_0$ 是 θ 的最小方差无偏估计（MVU 估计）。

3. 大样本性质

由于参数 θ 的估计 $\hat{\theta}$ 依赖于已有的样本值，设样本为 $X_i(i = 1, 2, \cdots, n)$，则我们可以令

$$\hat{\theta} = \hat{\theta}(X_1, X_2, \cdots, X_n)$$

下面研究 $\hat{\theta}$ 在 $n \to \infty$ 时的一些性质。

渐近无偏性

若当 $n \to \infty$ 时，估计的偏差趋于 0，也即成立

$$\lim_{n\to\infty} E_\theta(\hat{\theta} - \theta) = 0$$

则称 $\hat{\theta}$ 是 θ 的一个渐近无偏估计。

相合性

若 $\forall \varepsilon > 0$，均成立

$$\lim_{n\to\infty} P\left(\left|\hat{\theta} - \theta\right| \geq \varepsilon\right) = 0$$

则称 $\hat{\theta}$ 是 θ 的一个相合估计，也即

$$\hat{\theta} \xrightarrow{P} \theta$$

相合性是良好估计的自然要求，例如二阶中心矩 m_2 就是对总体方差 σ^2 的一个相合估计，因为

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\overline{X} - \mu)^2$$

而我们有

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \xrightarrow{P} \sigma^2$$

$$(\overline{X} - \mu)^2 \xrightarrow{P} 0$$

根据依概率收敛的四则运算法则，可以得到

$$m_2 \xrightarrow{P} \sigma^2$$

渐近正态性

若存在 σ_n，成立

$$\lim_{n\to\infty} \sigma_n = 0$$

$$\lim_{n\to\infty} P\left(\frac{\hat{\theta} - \theta}{\sigma_n} \leq x\right) = \Phi(x)$$

则称 $\hat{\theta}$ 是 θ 的一个相合渐近正态估计。例如，由 CLT 可得 $\overline{X}$ 是总体均值 μ 的一个相合渐近正态估计。

有时，我们可以取 $\sigma_n^2 = Var(\sigma)$。当 n 很大时，由上述定义，我们可以用二项分布 $N(\theta, \sigma_n^2)$ 来近似 $\hat{\theta}$ 的分布。

4. 极大似然估计与 Fisher 信息量

设 $X_i (i = 1, 2, \cdots, n)$ 为来自总体的独立同分布样本，参数为 θ，总体的分布为 $f(x;\theta)$，则有 MLE 过程中的似然函数

$$L(\theta) = f(X_1, X_2, \cdots, X_n; \theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

设极大似然估计值为 θ^*，极大似然估计满足渐近正态性，且若 f 满足光滑性条件，我们有

$$\exists \sigma_n, \frac{\theta^* - \theta}{\sigma_n} \longrightarrow N(0, 1)$$

下面引入 Fisher 信息量，用于求解符合条件的 σ_n。

对似然函数取自然对数，得到

$$l(\theta) = \sum_{i=1}^{n} \ln\left[f(X_i;\theta)\right]$$

将 $f(X_i;\theta)$ 简记为 f_i，$\forall i \in [1,n] \cap \mathbb{Z}$，我们有

$$\frac{\partial \ln f_i}{\partial \theta} = \frac{f_\theta}{f_i}$$

其中 f_θ 是总体分布对 θ 的边际分布。
求期望可得

$$E\left(\frac{\partial \ln f_i}{\partial \theta}\right) = \int \frac{f_\theta}{f_i} f_i \mathrm{d}x = \int f_\theta \mathrm{d}x = 0$$

则定义 Fisher 信息量（iid 表示独立同分布）：

$$\begin{aligned}
&I_n(\theta)\\
=&E\left[\left(\frac{\partial l(\theta)}{\partial \theta}\right)^2\right]\\
=&E\left[\left(\sum_{i=1}^{n}\frac{\partial \ln f_i}{\partial \theta}\right)^2\right]\\
=&\sum_{i=1}^{n} E\left[\left(\frac{\partial \ln f_i}{\partial \theta}\right)^2\right] + \sum_{i\neq j} E\left(\frac{\partial \ln f_i}{\partial \theta}\cdot\frac{\partial \ln f_j}{\partial \theta}\right)\\
\overset{\text{iid}}{=\!=}&\sum_{i=1}^{n} E\left[\left(\frac{\partial \ln f_i}{\partial \theta}\right)^2\right] + \sum_{i\neq j} E\left(\frac{\partial \ln f_i}{\partial \theta}\right)\cdot E\left(\frac{\partial \ln f_j}{\partial \theta}\right)\\
=&\sum_{i=1}^{n} E\left[\left(\frac{\partial \ln f_i}{\partial \theta}\right)^2\right]\\
=&nI(\theta)
\end{aligned}$$

其中

$$I(\theta) = E\left[\left(\frac{\partial \ln f_i}{\partial \theta}\right)^2\right], i \in [1,n] \cap \mathbb{Z}$$

由极大似然估计的定义，我们有

$$l'(\theta^*) = 0$$

利用 $l'(\theta)$ 的泰勒展开，我们有

$$l'(\theta^*) = l'(\theta) + (\theta^* - \theta)l''(\theta) + o(\theta^*)$$

近似地，我们可以得到

$$l'(\theta) + (\theta^* - \theta)l''(\theta) = 0$$

进一步得到

$$\sqrt{n}(\theta^* - \theta) = \frac{\frac{1}{\sqrt{n}}l'(\theta)}{-\frac{1}{n}l''(\theta)}$$

设随机变量 $Y_i = \dfrac{\partial \ln f_i}{\partial \theta}$，则有 $E(Y_i) = 0, Var(Y_i) = E(Y_i^2) - E^2(Y_i) = I(\theta)$，研究上述式子分子可得

$$\frac{1}{\sqrt{n}} l'(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i = \sqrt{n} \cdot \frac{1}{n} \sum_{i=1}^{n} Y_i = \frac{\overline{Y}}{\frac{1}{\sqrt{n}}} \xrightarrow{\text{LLN}} N(0, I(\theta))$$

研究分母可得

$$-\frac{1}{n} l''(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \ln f_i}{\partial \theta^2} \xrightarrow{P} -E\left(\frac{\partial^2 \ln f_i}{\partial \theta^2}\right)$$

又因为

$$\frac{\partial^2 \ln f_i}{\partial \theta^2} = \frac{f_{\theta\theta}}{f} - \left(\frac{\partial \ln f_i}{\partial \theta}\right)^2$$

则有

$$E\left(\frac{\partial^2 \ln f_i}{\partial \theta^2}\right) = -E\left[\left(\frac{\partial \ln f_i}{\partial \theta}\right)^2\right] = -E(Y_i^2) = -I(\theta)$$

于是

$$-\frac{1}{n} l''(\theta) \xrightarrow{P} I(\theta)$$

结合上述讨论可得

$$\sqrt{n}(\theta^* - \theta) \longrightarrow N\left(0, \frac{1}{I(\theta)}\right)$$

也即

$$\sqrt{nI(\theta)}(\theta^* - \theta) \longrightarrow N(0, 1)$$

于是可得

$$\sigma_n = \frac{1}{\sqrt{nI(\theta)}}$$

三、参数的 Bayes 估计

1. 先验分布的概念

在实际统计推断过程中，对于参数 θ，我们在搜集数据之前可能对其有一个先验认识，这个先验认识可以用一个概率分布来刻画，称为先验分布。用 Θ 表示 θ 分布的随机变量，则先验分布可以表示为 $f_\Theta(\theta)$。

2.Bayes 估计的过程

Bayes 估计是利用搜集的样本数据，将先验分布更新为后验分布的过程。设 X 代表试验过程的随机变量，则搜集的样本分布可以表示为 $f_X(x|\theta)$，则根据 Bayes 公式，我们可以得到 θ 的后验分布：

$$f_\Theta(\theta|x) = \frac{f(x,\theta)}{f_X(x)} = \frac{f_X(x|\theta) f_\Theta(\theta)}{f_X(x)}$$

其中

$$f_X(x) = \int_{\mathbb{R}} f(x,\theta) \mathrm{d}\theta$$

此时可以通过求后验分布期望或者后验分布众数的方式得到 Bayes 估计值。

3. 典例：抛硬币问题

一枚硬币抛了 n 次，出现了 x 次正面向上，估计正面向上的概率 θ。

根据同等无知原则，我们可以认为 θ 的先验分布 $\Theta \sim U(0,1)$，此时有 $f_\Theta(\theta)=1, \theta \in (0,1)$。

设随机变量 X 表示抛 n 次硬币正面向上的次数，则对给定的概率 θ，我们有 $X \sim B(n,\theta)$，于是我们有

$$f_X(x|\theta)=\binom{n}{x}\theta^x(1-\theta)^{n-x}, x \in [0,n] \cap \mathbb{Z}$$

于是我们有 θ 与 x 的联合分布：

$$f(x,\theta)=f_\Theta(\theta)f_X(x|\theta)=\binom{n}{x}\theta^x(1-\theta)^{n-x}$$

于是有

$$f_X(x)=\int_0^1 f(x,\theta)\mathrm{d}\theta=\frac{n!}{x!(n-x)!}\int_0^1 \theta^x(1-\theta)^{n-x}\mathrm{d}\theta=\frac{n!}{x!(n-x)!}\frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(n+2)}$$

其中

$$\Gamma(x)=\int_0^{+\infty} t^{x-1}e^{-t}\mathrm{d}t$$

由分部积分易得

$$\Gamma(x+1)=x\Gamma(x)$$

特别地，当 x 为正整数时，递归可以得到

$$\Gamma(x+1)=x!$$

于是我们有

$$f_X(x)=\frac{n!}{x!(n-x)!}\frac{x!(n-x)!}{(n+1)!}=\frac{1}{n+1}$$

于是我们更新得到 θ 的后验分布：

$$f_\Theta(\theta|x)=\frac{f(x,\theta)}{f_X(x)}=\frac{(n+1)!}{x!(n-x)!}\theta^x(1-\theta)^{n-x}=\frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)}\theta^x(1-\theta)^{n-x}$$

于是可得后验分布期望估计值为

$$\hat{\theta}=\int_0^1 \theta f_\Theta(\theta|x)\mathrm{d}\theta=\frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)}\int_0^1 \theta^{x+1}(1-\theta)^{n-x}\mathrm{d}x=\frac{x+1}{n+2}$$

后验分布众数即为使得后验分布 pdf 取最大值的 θ 值。函数 $f_\Theta(\theta|x)$ 对 θ 求导，令导函数值为 0，可以解得后验分布众数估计值为

$$\hat{\theta}=\frac{x}{n}$$

相关补充

1. 在本问题中，先验分布服从 Beta$(1,1)$，后验分布服从 Beta$(x+1,n-x+1)$。一般地，若先验分布服从 Beta(a,b)，则后验分布服从 Beta$(x+a,n-x+b)$。

2. 后验分布众数估计值与 MLE 估计值相等。这是因为在同等无知原则下，我们有

$$f_\Theta(\theta) \propto 1$$

于是有

$$f_\Theta(\theta|x) = \frac{f_\Theta(\theta) f(x|\theta)}{f_X(x)} \propto f(x|\theta) = f(x;\theta) = L(\theta)$$

则它们一定在同一点取极大值，从而后验分布众数估计值与 MLE 估计值相等。

四、参数的区间估计

有时我们难以准确地给出参数具体的估计值，此时我们可以在一定置信度下给出参数可能的取值区间，这就是参数的区间估计。

在下面出现的例子中，X 代表总体分布的随机变量，每个例子中都有 n 个独立同分布的已知样本 $X_i(i=1,2,\cdots,n)$，$\overline{X}$ 代表样本均值，S^2 代表样本方差。

1. 定义

给定 $\alpha \in (0,1)$，若对参数 θ 的任意可能取值，都成立

$$P\left(\hat{\theta}_1 < \theta < \hat{\theta}_2\right) \geq 1-\alpha$$

则称 $\left(\hat{\theta}_1, \hat{\theta}_2\right)$ 为 θ 的 $(1-\alpha)$ 置信的区间估计。

2. 典例

1. 设 $X \sim N(\mu, \sigma^2)$，σ^2 已知，μ 未知，给出 μ 的 $(1-\alpha)$ 置信区间估计。

我们熟知

$$\overline{X} - \mu \sim N\left(0, \frac{\sigma^2}{n}\right)$$

于是我们有

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

则根据标准正态分布的 pdf 图像，我们有

$$-z_{\frac{\alpha}{2}} < \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < z_{\frac{\alpha}{2}}$$

其中 $z_{\frac{\alpha}{2}}$ 为 $N(0,1)$ 的上侧 $\dfrac{\alpha}{2}$ 分位数。

于是我们可得所求的估计区间为

$$\left(\overline{X}-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\overline{X}+z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

2. 设 $X\sim N(\mu,\sigma^2)$，μ,σ^2 未知，给出 μ 的 $(1-\alpha)$ 置信区间估计。

   我们有

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

$$\frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$$

   则有

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

   根据 t 分布的图像，我们有

$$-t_{\frac{\alpha}{2}}(n-1)<\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}}<t_{\frac{\alpha}{2}}(n-1)$$

   其中 $t_{\frac{\alpha}{2}}(n-1)$ 为自由度 $n-1$ 的 t 分布的上侧 $\dfrac{\alpha}{2}$ 分位数。

   于是我们可得所求的估计区间为

$$\left(\overline{X}-t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}},\overline{X}+t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}},\right)$$

补充

由上述过程可见，在进行区间估计时，需要利用样本统计量，选择合适的分布，避开未知的参数，仅留下需要估计的参数，之后利用分布的分位数来给出具体的估计区间。

3. 枢轴变量法

步骤

1. 找出参数 θ 的相关统计量 $\hat{\theta}(X_1,X_2,\cdots,X_n)$。
2. 找出枢轴变量 $H(\hat{\theta},\theta)$ 的分布，需要与未知变量无关。
3. 求出枢轴变量分布的对应分位数，之后给出所要求的估计区间。

典例

1. 设总体期望为 μ，方差为 σ^2，μ,σ^2 均未知，给出 σ^2 的 $(1-\alpha)$ 置信区间估计。

   此时我们有枢轴变量

$$\frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$$

于是可以得到所求估计区间为

$$\left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}, \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}\right)$$

2. 设 $X \sim N(\mu_1, \sigma^2)$，$Y \sim N(\mu_2, \sigma^2)$，$X$ 与 Y 独立，μ_1, μ_2, σ^2 未知，给出 $\mu_1 - \mu_2$ 的 $(1-\alpha)$ 置信区间估计。

   我们设 $X_1, X_2, \cdots, X_n, Y_1, Y_2, \cdots, Y_m$ 为随机样本，S_1^2 为 X 的样本方差，S_2^2 为 Y 的样本方差。则我们有

$$(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2) \sim N\left(0, \frac{\sigma^2}{n} + \frac{\sigma^2}{m}\right)$$

   我们作如下代换：

$$\sigma'^2 := \frac{\sigma^2}{n} + \frac{\sigma^2}{m}$$

$$S^2 := \frac{n-1}{n+m-2}S_1^2 + \frac{m-1}{n+m-2}S_2^2$$

   因为

$$\frac{(n-1)S_1^2}{\sigma^2} + \frac{(m-1)S_2^2}{\sigma^2} \sim \chi^2(n+m-2)$$

   所以有

$$\frac{(n+m-2)S^2}{\sigma^2} \sim \chi^2(n+m-2)$$

   于是我们有

$$\frac{\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sigma'}}{\frac{S}{\sigma}} \sim t(n+m-2)$$

   也即

$$\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t(n+m-2)$$

   上面就是我们估计所用的枢轴变量。于是可以得到所求估计区间为

$$\left(\overline{X} - \overline{Y} - t_{\frac{\alpha}{2}}(n+m-2)S\sqrt{\frac{1}{n} + \frac{1}{m}}, \overline{X} - \overline{Y} + t_{\frac{\alpha}{2}}(n+m-2)S\sqrt{\frac{1}{n} + \frac{1}{m}}\right)$$

补充：关于大样本方法

当样本数 n 较大时，一般采用渐近置信区间估计，依赖于中心极限定理等方法，找出参数的近似分布进而解决问题，例如可以用样本方差 S^2 或二阶中心矩 m_2 来近似 σ^2。采用这种方法给出的也是渐近置信区间。

4.Bayes 区间估计

定义

假设根据 Bayes 估计过程, 已经得到了参数 θ 的后验分布 $f_{\Theta}(\theta|x)$, 给定观测值 x 和实数 $\alpha \in (0,1)$, 若对参数 θ 的任意可能取值，都成立

$$P\left(a < \Theta < b|x\right) \geq 1-\alpha$$

则称 (a,b) 为 θ 的 $(1-\alpha)$ 置信的 Bayes 区间估计。

典例

设 $X \sim N(\mu, \sigma^2)$，σ^2 已知，给出 μ 的 $(1-\alpha)$ 置信的 Bayes 区间估计。

我们取 μ 的先验分布 $f(\mu) \propto 1$，则我们有后验分布为 $N\left(\overline{X}, \dfrac{\sigma^2}{n}\right)$。于是我们进一步可得

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

于是可以得到所求 Bayes 估计区间为

$$\left(\overline{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

五、假设检验

在实际研究问题时，对于总体的某些未知性质，我们往往会提出一些假设。搜集到样本数据之后，我们需要研究样本在多大程度上支持之前提出的假设。这样的过程就是假设检验的过程。

1. 检验的基本概念

假设的基本定义

1. 统计假设：对一个或多个总体的某种推断或者猜测。
2. 原假设（null hypothesis）：一般用 H_0 表示，代表被检验的假设。
3. 备择假设（alternative hypothesis）：也叫研究假设，一般用 H_1 表示，代表拒绝 H_0 后可供选择的假设。

假设的参数表示

设参数 θ 的可能取值集合为 Ω，若假设可以被表示成参数形式，那么原假设和备择假设可以表示为：

$$H_0 : \theta \in \Theta_0, H_1 : \theta \in \Theta_1$$

其中

$$\Theta_0 \cap \Theta_1 = \varnothing, \Theta_0 \cup \Theta_1 = \Omega$$

根据检验的不同类别，参数形式表示也有不同的方式。例如，我们估计某个总体的均值 μ，原假设 $H_0:\mu=\mu_0$，备择假设为 $H_1:\mu\neq\mu_0$，这是双边检验；我们研究两个总体的均值 μ_1 和 μ_2 之间的关系，原假设为 $H_0:\mu_1=\mu_2$，备择假设为 $H_1:\mu_1>\mu_2$，这是单边检验。

检验的概念与抽象表示

假设检验是一个依据样本的决策过程，决定所观测的样本值是否拒绝 H_0。对于所有可能观测组成的集合 $\Omega=\{(X_1,X_2,\cdots,X_n)\}$，我们定义临界域 R：

$$R=\{(X_1,X_2,\cdots,X_n)|T(X_1,X_2,\cdots,X_n)\geq c\}$$

其中 $T(X_1,X_2,\cdots,X_n)$ 表示了样本的某个统计量，c 是所选取的临界值。当我们所观测到的样本属于临界域 R 时，则拒绝 H_0；若属于临界域对 Ω 的补集 R^c，则不拒绝 H_0。

检验的基本原则是，在假设 H_0 为真的情况下，若所观测到的样本的理论出现概率很小，则意味着样本拒绝 H_0。这是依据小概率事件不容易发生的事实作出的判断。

检验时两种可能的错误

1. I 类错误——弃真错误：原假设 H_0 为真时拒绝 H_0。
2. II 类错误——纳伪错误：原假设 H_0 为假时不拒绝 H_0。

功效函数

假设给定临界域 R，则发生 I 类错误的概率为

$$P(\mathrm{I})=P_\theta((X_1,X_2,\cdots,X_n)\in R),\text{when }\theta_0\in\Theta_0$$

记为 $\alpha(R)$。发生 II 类错误的概率为

$$P(\mathrm{II})=P_\theta((X_1,X_2,\cdots,X_n)\in R^c),\text{when }\theta_1\in\Theta_1$$

记为 $\beta(R)$。

功效函数给出了对不同给定的 θ，给定临界域 R，拒绝原假设 H_0 的概率。借助上面的定义，我们可以给出功效函数的定义：

$$P_\theta((X_1,X_2,\cdots,X_n)\in R)=\begin{cases}\alpha(R), & \theta\in\Theta_0\\ 1-\beta(R), & \theta\in\Theta_1\end{cases}$$

Neyman-Pearson 范式

这是进行假设检验时，对于给定检验水平 α 和固定的样本容量 n，常用的划分临界域 R 的方法。其核心思想就是控制 $P(\mathrm{I})\leq\alpha$，再在这一限制下使得 $P(\mathrm{II})$ 尽可能小。

其他补充

1. 原假设 H_0 和备择假设 H_1 一般是地位不对等的。原假设 H_0 通常是受到保护的，检验过程是在试图从已观测样本中找到证据来拒绝 H_0。而备择假设通常是在研究中感兴趣的。

2. 在检验水平 α 固定的情况下，使得 $P(\mathrm{II})$ 最小的检验成为水平 α 下的一致最优检验。其不一定存在，一般也不易求解。

2. 临界值检验

步骤

1. 提出原假设 H_0 和备择假设 H_1。

2. 给定检验水平 $\alpha > 0$。

3. 确定检验过程中需要用到的关于样本的统计量，例如样本均值或样本方差，确定拒绝域 R 的形状（根据 H_1 判断此时的检验是单边检验还是双边检验，从而判定需要单边拒绝还是双边拒绝）。

4. 根据 Neyman-Pearson 范式，由 $P(\mathrm{I}) \leq \alpha$，给出拒绝域 R。

5. 采样得到所检验统计量的值。

6. 根据划定的拒绝域 R 进行决策。

典例

1. 设总体均值为 μ，方差为 σ^2，σ^2 已知。现提出原假设 $H_0 : \mu = \mu_0$，备择假设 $H_1 : \mu \neq \mu_0$，给定检验水平 α 和已观测样本 $X_i (i = 1, 2, \cdots, n)$，对原假设进行检验。

   根据 H_1 的表述，本题的拒绝域为双边拒绝，基本思路是当 $\left|\overline{X} - \mu_0\right| \geq c$ 时就拒绝，c 为划分拒绝域的参数，也即所谓的临界值。根据 Neyman-Pearson 范式，我们控制

   $$P(\mathrm{I}) = P\left(\left|\overline{X} - \mu_0\right| \geq c\right) \leq \alpha$$

   也即

   $$P\left(\left|\frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right| \geq \frac{c}{\frac{\sigma}{\sqrt{n}}}\right) \leq \alpha$$

   根据 CLT，我们有

   $$\frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

   于是我们得到此时的检验准则：在 α 的检验水平下拒绝 H_0 的条件为

   $$\left|\overline{X} - \mu_0\right| \geq z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

   其中 $\overline{X}$ 是已观测样本的均值，$z_{\frac{\alpha}{2}}$ 是 $N(0, 1)$ 的上 $\dfrac{\alpha}{2}$ 分位数。

2. 设总体均值为 μ，方差为 σ^2，σ^2 已知。现提出原假设 $H_0:\mu\geq\mu_0$，备择假设 $H_1:\mu<\mu_0$，给定检验水平 α 和已观测样本 $X_i(i=1,2,\cdots,n)$，对原假设进行检验。

   根据 H_1 的表述，本题的拒绝域为单边拒绝，基本思路为当 $\overline{X}<c$ 时就拒绝，c 为划分拒绝域的参数，也即所谓的临界值。则根据 Neyman-Pearson 范式，当 H_0 为真时，我们控制

$$P\left(\overline{X}\leq c\right)\leq\alpha$$

   则有

$$P\left(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\leq\frac{c-\mu}{\frac{\sigma}{\sqrt{n}}}\right)\leq\alpha$$

   由于此时有

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

   所以有

$$\Phi\left(\frac{c-\mu}{\frac{\sigma}{\sqrt{n}}}\right)\leq\alpha$$

   对一切 $\mu\geq\mu_0$ 都成立，其中 $\Phi(x)$ 是标准正态分布的 cdf。于是我们有

$$\frac{c-\mu_0}{\frac{\sigma}{\sqrt{n}}}\leq-z_\alpha$$

   于是我们得到此时的检验准则：在 α 的检验水平下拒绝 H_0 的条件为

$$\overline{X}\leq\mu_0-z_\alpha\frac{\sigma}{\sqrt{n}}$$

   其中 $\overline{X}$ 是已观测样本的均值，z_α 是 $N(0,1)$ 的上 α 分位数。

临界值检验与区间估计的关系

设总体分布为 $N(\mu,\sigma^2)$，σ^2 已知，给定 $\alpha\in(0,1)$，$X_i(i=1,2,\cdots,n)$ 为独立同分布的样本。则我们知道 μ 的 $(1-\alpha)$ 置信区间为

$$\left(\overline{X}-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\overline{X}+z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

我们考虑以 α 为检验水平对 μ 进行假设检验，提出原假设 $H_0:\mu=\mu_0$，$H_1:\mu\neq\mu_0$，则根据已有讨论，我们得到拒绝 H_0 的条件为

$$\left|\overline{X}-\mu_0\right|\geq z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

则此时拒绝域的补集为

$$R^c=\left\{(X_1,X_2,\cdots,X_n)\middle|\left|\overline{X}-\mu_0\right|<z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right\}$$

也即

$$\mu_0\in\left(\overline{X}-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\overline{X}+z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

由此可见，$\mu_0\in$ 置信区间 $\Leftrightarrow$ 假设检验（$H_0:\mu=\mu_0,H_1:\mu\neq\mu_0$）不拒绝 H_0，拥有良好的对偶关系。

3.p 值检验

p 值的定义

p 值是指当原假设 H_0 为真时，观测值以更极端的情况出现的概率。所谓的更极端情况是由备择假设 H_1 决定的。一般而言，关于 p 值的计算我们有如下定理：

设 $X_i(i = 1, 2, \cdots, n)$ 表示某一个可能出现的独立同分布的 n 个样本，$x_i(i = 1, 2, \cdots, n)$ 表示已观测的 n 个独立同分布的样本。若拒绝 H_0，$\theta \in \Theta_0$ 等价于 $T(X_1, X_2, \cdots, X_n) \geq c$，其中 $T(X_1, X_2, \cdots, X_n)$ 代表了某个统计量，则 p 值等于 $\sup\limits_{\theta \in \Theta_0} P\left(T(X_1, X_2, \cdots, X_n) \geq T(x_1, x_2, \cdots, x_n)\right)$。

步骤

1. 提出原假设 H_0 和备择假设 H_1。
2. 给定检验水平 $\alpha > 0$。
3. 确定检验过程中需要用到的关于样本的统计量。
4. 采样得到所检验统计量的值。
5. 根据定义计算 p 值。
6. 根据 p 值和检验水平的大小关系进行决策。

典例

此处以选举问题作为 p 值检验的例子。设调查到的选举支持率为 p_n，样本容量为 n。设真正的支持率为 p。提出原假设 $H_0 : p = p_0$，备择假设 $H_1 : p > p_0$，则当 H_0 为真时，考虑某个可能出现的支持率调查结果 P_n，根据 CLT，我们近似认为

$$\frac{P_n - p_0}{\hat{se}(P_n)} \sim N(0, 1)$$

其中样本标准差 $\hat{se}(P_n) = \sqrt{\dfrac{p_n(1-p_n)}{n}}$ 或 $\sqrt{\dfrac{p_0(1-p_0)}{n}}$。根据备择假设的描述，此时的 p 值为

$$P\left(\frac{P_n - p_0}{\hat{se}(P_n)} \geq \frac{p_n - p_0}{\hat{se}(P_n)}\right) = 1 - \Phi\left(\frac{p_n - p_0}{\hat{se}(P_n)}\right)$$

4.Bayes 假设检验

设原假设为 H_0，备择假设为 H_1，事件 X 表示已观测的样本值出现，Bayes 检验的过程中需要研究后验概率，拒绝 H_0 的条件为

$$\frac{P(H_0|X)}{P(H_1|X)} = \frac{P(H_0)P(X|H_0)}{P(H_1)P(X|H_1)} < c$$

其中 c 为人为设定的阈值，例如可以设定 $c = 1$。

5. 拟合优度检验

拟合优度检验一般用于检验某些事件的发生是否服从某个特定的分布。

设所观测的事件为 $X_i(i=1,2,\cdots,k)$，X_i 发生的概率为 p_i，提出原假设 $H_0: P(X=a_i)=p_i(i=1,2,\cdots,k)$，也即假设这 k 个事件服从某个特定分布。进行 n 次观测，考虑 χ_0^2 统计量：

$$\chi_0^2=\sum_{i=1}^{k}\frac{(O_i-E_i)^2}{E_i}$$

其中 O_i 为事件 X_i 的观测频数，E_i 为事件 X_i 的期望频数（也即 np_i）。

我们认为若 H_0 为真，则当 $n\to\infty$ 时，$\chi_0^2\to\chi^2(k-1)$。在这样的近似下，我们就可以计算对应的 p 值进行检验，此时的 p 值为 $P_{H_0}(M\geq\chi_0^2)$，其中 M 为服从自由度 $(k-1)$ 的卡方分布的随机变量。

在实际应用中，我们至少需要确保每个 $E_i\geq 5$，才能较好地运用这样的近似方法。因此，在实际问题中我们常常选择将 $E_i<5$ 的事件进行合并考虑，结合极大似然估计等方法计算得到合并后新的 E_i 值，并且进一步计算得到合并后的 χ_0^2 统计量。设合并后的事件数为 k，合并后的参数维数为 s，则合并后检验 p 值的上界为 $\chi^2(k-1-s)$ 对应的 p 值，下界为 $\chi^2(k-1)$ 对应的 p 值。

拟合优度检验也可以用于数据可信度/数据造假相关问题的检验。

6. 列联表检验

列联表检验常用于独立性检验，以下通过一个例子来介绍列联表检验的步骤：检验饮酒口味偏好与饮酒者性别是否独立。

在本例中，收集到的数据如下表：

	清淡	普通	浓重
男	20	40	20
女	30	30	10

在列联表检验中，设搜集的数据有 a 行 b 列，对第 i 行第 j 列的元素赋以一个概率 p_{ij}，第 i 行的行概率为 p_{i+}，第 j 列的列概率为 p_{+j}，则提出原假设为 $H_0: p_{ij}=p_{i+}P_{+j}$，这就表现了数据之间的独立性。设所检验的总人数为 n，检验过程中所使用的统计量为

$$\chi_0^2=\sum_{i,j}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

其中 O_{ij} 为第 i 行第 j 列的观测值，$E_{ij}=np_{ij}$，为期望值。

当 H_0 为真时，我们使用极大似然估计法来近似计算 p_{ij}，则有

$$p_{ij}^*=p_{i+}^*p_{+j}^*=\frac{\sum_j O_{ij}}{n}\cdot\frac{\sum_i O_{ij}}{n}$$

于是我们有

$$E_{ij}=np_{ij}\approx np_{ij}^*=\frac{(\sum_i O_{ij})(\sum_j O_{ij})}{n}$$

这样我们就可以计算得到 χ_0^2 统计量，其分布可以近似看作自由度为 $(a-1)(b-1)$ 的卡方分布，于是可以进一步计算得到 p 值，从而作出决策。

例如，在本例中，计算得到 $\chi_0^2\approx 6.12$，p 值为 $P(\chi^2(2)\geq\chi_0^2)\approx 0.047$，p 值很小，因此不拒绝 H_0，认为独立性成立。

7. 似然比检验

定义

考虑 Bayes 假设检验的过程，采用的后验概率之比为

$$\frac{P(H_0|X)}{P(H_1|X)}=\frac{P(H_0)P(X|H_0)}{P(H_1)P(X|H_1)}$$

我们将 $\dfrac{P(H_0)}{P(H_1)}$ 称为先验比，$\dfrac{P(X|H_0)}{P(X|H_1)}$ 称为似然比。在不考虑先验比的情况下，利用似然比进行检验，也即拒绝 H_0 的条件为

$$\frac{P(X|H_0)}{P(X|H_1)}<c$$

其中常数 c 由给出的检验水平 α 决定，

上述过程即为似然比检验的过程。当假设都是简单假设时，似然比检验就是最优检验（功效最大的检验）。

广义似然比检验

设原假设为 $H_0:\theta\in\Theta_0$，备择假设为 $H_1:\theta\in\Theta_1$，样本彼此独立同分布。则定义广义似然比为

$$\Lambda^*:=\frac{\sup\limits_{\theta\in\Theta_0}L(\theta)}{\sup\limits_{\theta\in\Theta_1}L(\theta)}$$

其中 $L(\theta)$ 为极大似然估计中使用的似然函数。事实上，由于技术原因，我们无法直接使用 Λ^*，因此考虑采用检验统计量 Λ 代替：

$$\Lambda:=\frac{\sup\limits_{\theta\in\Theta_0}L(\theta)}{\sup\limits_{\theta\in\Theta_1\cup\Theta_0}L(\theta)}=\min\{\Lambda^*,1\}$$

此时我们有

$$\sup_{\theta\in\Theta_1\cup\Theta_0}L(\theta)=L(\theta^*)$$

其中 θ^* 为极大似然估计值，较为方便计算。

在广义似然比检验中，Λ 越小，则样本越反对 H_0，由此可以作出决策。

典例：多项分布

设我们对多项分布提出假设 $H_0:p_1=p_{10},p_2=p_{20},\cdots,p_k=p_{k0}$，总观测数为 n，每个单元的观测频数为 $n_1,n_2,\cdots,n_k$，其中 $\sum\limits_{i=1}^{k}p_i=n,\sum\limits_{i=1}^{k}n_i=n$。则我们得到似然函数为

$$L(p_1,p_2,\cdots,p_k)=\binom{n}{n_1,n_2,\cdots,n_k}\prod_{i=1}^{k}p_i^{n_i}$$

利用广义似然比检验，我们有

$$\Lambda=\frac{L(p_1,p_2,\cdots,p_k)}{L(p_1^*,p_2^*,\cdots,p_k^*)},p_{i^*}=\frac{n_i}{n}$$

设 O_i 表示观测频数，$E_i = np_i$ 表示期望频数，于是我们有

$$\begin{aligned}
&-2\ln\Lambda \\
=&-2\left(\sum_{i=1}^{k} n_i \ln\frac{p_i}{p_i^*}\right) \\
=&2\left(\sum_{i=1}^{k} O_i \ln\frac{np_i^*}{np_i}\right) \\
=&2\left(\sum_{i=1}^{k} O_i \ln\frac{O_i}{E_i}\right) \\
\xlongequal{\text{Taylor}}&2\sum_{i=1}^{k}(O_i - E_i) + \sum_{i=1}^{k}\frac{(O_i-E_i)^2}{E_i} + o((O_i-E_i)^2) \\
\approx&\sum_{i=1}^{k}\frac{(O_i-E_i)^2}{E_i}
\end{aligned}$$

由此可见，$\dim(H_0) = 0$，没有自由度；$\dim(\Theta_0 \cup \Theta_1) = k-1$，自由度为 $(k-1)$。于是我们可以通过近似为自由度 $(k-1)$ 的卡方分布的方法来计算 p 值，从而做出决策。

结语

磨磨蹭蹭几个星期，终于把后半学期统计推断部分的知识点整理完了。不同于上学期概率论部分的内容，统计推断部分的一些知识较为抽象，需要结合更多的具体例子进行理解，因此在编写时可能有所疏漏与错误，敬请谅解，或可以通过邮箱：2322751077@qq.com 进行联系。

希望本文档可以对大家的概率论与数理统计课程学习起到一些辅助的作用，这也是我在寒假编写这样一份文档的初衷。