假设检验（下）

拟合优度检验

推导

前面的假设检验都假定了总体分布的形式是已知的，本节介绍总体分布的检验方法。设 $X_1, X_2, \cdots, X_n$ 是总体 X 的样本。对于已知的概率分布函数 $F(x)$，考虑假设

$$H_0: X \sim F(x) \text{ vs } H_1: X \nsim F(x)$$

的检验问题。$X \sim F(x)$ 表示 X以$F(x)$ 为分布函数。

拟合优度检验考虑的是观测样本及其总体分布是否能拟合，以及拟合好坏的标准。给定总体 X 的样本观测值 $X_1, X_2, \cdots, X_n$，取$t_0 < \min\{X_1, X_2, \cdots, X_n\}$，$t_m > \max\{X_1, X_2, \cdots, X_n\}$，类似于制作频率直方图的方法，取 $t_0 < t_1 < t_2 < \cdots < t_m$，然后将区间 $(t_0, t_m]$ 进行划分，得到互不相交的区间$l_j = (t_{j-1}, t_j], j = 1, 2, \cdots, m$。

下面用观测样本落入区间 I_j 的频率$\hat{p}_j = \frac{1}{n}\sum_{k=1}^{n} \mathrm{I}[X_k \in I_j]$，作为概率$p_j = P(X \in I_j) = F(t_j) - F(t_{j-1})$的估计。

用$U = \sum_{j=1}^{m} \frac{n}{p_j}\left(\hat{p}_j - p_j\right)^2$描述频率 $\{\hat{p}_j\}$ 和概率 $\{p_j\}$ 之间的差异。对于较大的样本量 n，在$H_0: X \sim F$下，从频率和概率的关系知道 $\left(\hat{p}_j - p_j\right)^2$ 应当较小，所以当 U 较大时应当拒绝 H_0。

在 H_0 下可以证明：当 n 较大时，U 近似服从 $m-1$ 个自由度的 χ^2 分布。于是 $H_0: X \sim F(x)$ 的显著水平（近似）为 α 的拒绝域是$W = \left\{U > \chi^2_\alpha(m-1)\right\}$。

> 如果总体分布 $F(x)$ 中有 r 个未知参数，就需要**用观测数据先计算出这 r 个未知参数的最大似然估计**，用最大似然估计代替真实参数后才能计算出 p_j。这时在 H_0 下可以证明：当 n 较大时，U 近似服从 $m-r-1$ 个自由度的 χ^2 分布。于是 H_0 的显著水平（近似）为 α 的拒绝域是$W = \left\{U > \chi^2_\alpha(m-r-1)\right\}$。
>
> 实际应用中，为了使得近似的程度较好，还应当要求样本量的大小和区间的划分满足以下的条件$np_j \geqslant 5, 1 \leqslant j \leqslant m$。**如果有不满足$np_j \geq 5$的，需要将区间合并！**

事例

自 1500-1931年的 $N = 432$ 年间，比较重要的战争在全世界共发生了299次。以每年为一个时间段的记录如下：

爆发的战争数 k	爆发 k 次战争的年数 m_k	频率 m_k/N	$P(Y=k)$
0	223	0.516	0.502
1	142	0.329	0.346
2	48	0.111	0.119
3	15	0.035	0.028
4+	4	0.009	0.005
总计	432	1.000	1.000

表中 $Y \sim \mathcal{P}(0.69), 0.69 = 299/432$ 是平均每年爆发的战争数。

用 X_j 表示第 j 年的战争数，则在 $H_0: X \sim \mathcal{P}(\lambda)$ 下，$X_1, X_2, \ldots, X_n$ 是泊松总体 $\mathcal{P}(\lambda)$ 的样本，其中未知参数 λ 的最大似然估计是 $\hat{\lambda} = \bar{X}_n = 299/432 = 0.69$，这时 H_0 中的 $\mathcal{P}(\lambda) = \mathcal{P}(0.69)$。将 $[0, \infty)$ 划分成 $m = 4$ 段：

$$l_1 = [0, 0.5], l_2 = (0.5, 1.5], l_3 = (1.5, 2.5], l_4 = (2.5, \infty)$$

分别计算出

$$\begin{aligned}
p_1 &= e^{-\hat{\lambda}} = 0.502, & \hat{p}_1 &= \tfrac{223}{432} = 0.516 \\
p_2 &= \hat{\lambda}e^{-\hat{\lambda}} = 0.346, & \hat{p}_2 &= \tfrac{142}{432} = 0.329 \\
p_3 &= \tfrac{\hat{\lambda}^2}{2!}e^{-\hat{\lambda}} = 0.119, & \hat{p}_3 &= \tfrac{48}{432} = 0.111 \\
p_4 &= 1 - \textstyle\sum_{j=1}^{3} p_j = 0.033, & \hat{p}_4 &= \tfrac{19}{432} = 0.044.
\end{aligned}$$

因为 $432 \times 0.033 = 14.256 > 5$，所以拟合条件成立，计算出

$$U = \sum_{j=1}^{4} \frac{n}{p_j}\left(\hat{p}_j - p_j\right)^2 = 2.3458$$

自由度为 $4-1-1=2$，查表得到

$$\chi^2_{0.05}(2) = 5.991 > U = 2.3458$$

所以不能拒绝总体 X 服从泊松分布 $\mathcal{P}(0.69)$。由于 n 较大，所以可以接受 H_0，接受 $X \sim \mathcal{P}(0.69)$。以 $\{U \geqslant 2.3458\}$ 作拒绝域，可以计算出拒绝 H_0 犯错误的概率

$$P = P\left(\chi^2_2 \geqslant 2.3458\right) = 0.3095$$

其中 χ_2^2 是服从 $\chi^2(2)$ 分布的随机变量。$P=0.3095$ 又称为**拟合优度**，它显示了数据和泊松分布 $\mathcal{P}(0.69)$ 的拟合情况。拟合优度越大，数据和假设分布的拟合程度越好。

列联表独立性检验（看PPT）