Homework Week3

20199102-Sun Qilong

March 13, 2020

Problem 1 (设计 χ^2 检验项目)

1. 从 $0, 1, ..., 2n-1$ 中随机取一个数，取完后放回。独立重复这个过程 k 次, 求事件 A_m: “此 k 个数中最大的数是 m” 的概率 $p_m(m \leq 2^n - 1)$。(提示：$p_m = \frac{(m+1)^k - m^k}{2^{nk}}$)

2. 根据上面的概率计算，设计一个随机检验项目。要求：

   (a) 选定参数 n 和 k。

   (b) 写出检测算法过程，包括观察值和计算过程

   (c) 说明确定 P 值的过程 (借助不完全 Γ 函数)。

answer

1. 计算分如下四步：

   (a) 先考虑所有的可能性。数字的总数为 2^n，有放回的取数，每次取得每个数字的个数相同。共有 2^{nk} 种可能性。

   (b) 考虑 k 次取数结果均不大于 m 的可能性。每次可取的数字总数为 m，有放回的取，每次取得每个数字的可能性相同。共有 m^k 种可能性。

   (c) 考虑 k 次取数结果均不大于 $m-1$ 的可能性。每次可取的数字总数为 $m-1$，有放回的取，每次取得每个数字的可能性相同。共有 $m-1^k$ 种可能性。

   (d) 满足题目条件的结果为 k 次结果均不大于于 m，但不能均小于 m。故可行解共有 $m^k - m - 1^k$ 种。故 $p_m = \frac{(m+1)^k - m^k}{2^{nk}}$

2. 随机检验项目具体步骤如下。

   (a) 选定参数 n 和 k, 其中 n 代表每相邻 n 位二进制组合形成一个数，k 代表相邻 k 个数作为一组进行下列的计算。

   (b) 考虑到在随机的情况下每个位置 0　1 的可能性相同，则相邻 n 位二进制组成的数字的可能也是服从 $U(0, 2^n - 1)$ 的分布。故每个大小为 k 的分组中，不同数字的分布服从 (1.) 中的结果。

   (c) 将 2^n 个不同的结果对应的概率 p_m 分成 t 个统计段 $e_1, e_2, ..., e_t$，使得每段的概率总和近似等于 $\frac{1}{t}$。这里我们假设实际每一段的概率为 ρ_i $(1 \leq i \leq t)$

   (d) 对于给定的二进制数据，每 nk 划分为一组，假设共计 N 组。统计每一组中最大值落入 $e_1, e_2, ..., e_t$ 的频数，假设为 $\eta_1, \eta_2, ..., \eta_t$。

(e) 构造统计量 $\xi = \sum_{i=1}^{t} \frac{\eta_i - N*\rho_i}{N*\rho_i}$, 则统计量应符合分布 $\chi^2(N-1)$

(f) 关于 P 值的计算。计算统计量 ξ_{obs} 公式如上文所述，则 $p_{value} = igamc(\frac{N-1}{2}, \frac{\xi_{obs}}{2})$

(g) 若 $p_{value} > P$, 则接受假设。

Problem 2 (分布距离)

1. 证明样本$1, 2, ..., n$ 上的概率分布 p 与均匀分布 u 之间的统计距离 $d(n,p) \leq \frac{n-1}{n}$。

2. 问题 (1.) 中的等号能达到吗？在那个概率分布上达到？

answer

1. 假设 P_i 代表样本$1, 2, ..., n$ 上的概率分布 P 中 $P(x=i)$ 的取值。并且已知均匀分布 $U(1,n)$ 中 $P(x=i) = \frac{1}{n}$。并且我们假设 $i = i_1, i_2, ..., i_k$ 时 $P_i > \frac{1}{n}$; 同时 $j = j_1, j_2, ..., j_{n-k}$ 时 $P_j \leq \frac{1}{n}$。

   当 $k = 0$ 时,$P \sim U(1,n)$，$d(n,p) = 0$。下考虑 $1 \leq k$

$$
\begin{aligned}
d(n,p) &= \frac{1}{2}[\sum_{t=1}^{k}(P_{i_t} - \frac{1}{n}) + \sum_{t=1}^{n-k}(\frac{1}{n} - P_{j_t})] \\
&= \frac{1}{2}(\sum_{t=1}^{k} P_{i_t} - \sum_{t=1}^{n-k} P_{j_t} - \frac{k}{n} + \frac{n-k}{n}) \\
&= \frac{1}{2}(1 - 2\sum_{t=1}^{n-k} P_{j_t} + \frac{n-2k}{n}) \\
&= \frac{1}{2}(2 - 2\sum_{t=1}^{n-k} P_{j_t} - \frac{2k}{n}) \\
&= 1 - \frac{k}{n} - \sum_{t=1}^{n-k} P_{j_t} \\
&\leq \frac{n-1}{n} - \sum_{t=1}^{n-k} P_{j_t} \qquad (1) \\
&\leq \frac{n-1}{n} \qquad (2)
\end{aligned}
$$

2. 考虑上式中 (1), (2) 中取等的条件 $k = 1$ 和 $\sum_{t=1}^{n-k} P_{j_t} = 0$。考虑概率的非负性, 则 $P_{j_t} = 0 (1 \leq t \leq n-1)$。故可得，$d(n,p) \leq \frac{n-1}{n}$ 的取等条件为，P 代表的是分布$P_i = 0 (i \neq s), P_s = 1 (\forall s \in [1,n])$