第三章 选择最大信息预报因子

章节引入

早在天气图出现之前，民间就已经广泛流传着有关天气的谚语。因为天气与人类的生活是密切相关的。谚语所反映的就是前期的征兆与后期天气的统计关联性。仅就降水而言，降水过程有共同性，也有特殊性，**单一预报指标**必然不能用在所有的降水过程。选择**最大信息的预报指标**（因子）可以减少天气预报的漏报率和空报率。

3.1 概率和条件概率以及预报指标

3.1.1 状态要素与随机事件

状态要素 指表征气象要素的各种状态，观测结果用**非数值型**数据表征。

事件 自然界中的各种现象。状态要素的各种状态可视为**随机事件**，例如：下雨、不下雨、小雨、大雨等都可能出或不出现，每一种状态都是随机的。

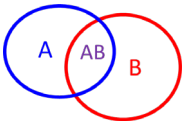
3.1.2 频率与概率

频率 衡量事件出现**可能性大小**的数量指标。n次观测次数中，事件A出现n_A次，则事件A的频率为 n_A/n。

概率 当观测的资料n足概大时（接近总体），$P(A)$ **稳定接近某个常数**时，这就是概率。

概率是事件的**总体特征**（频率的理论值），**频率**是事件的**样本值**（概率的估计值）。

3.2 条件概率和天气预报指标



条件概率 在事件 B 已经发生的条件下计算事件 A 的概率(此处都是状态要素，非连续)，称为事件 A 在事件 B 已出现条件下的条件概率，记为$P(A/B)$。若事件 AB 同时出现的概率为$P(AB)$，则 $\boldsymbol{P(A/B)=\frac{P(AB)}{P(B)}}$

重要意义 **条件概率是统计预报的基础**。统计天气预报中，往往将事件 A 取为所要预报的具体内容，而将 B 取为事件 A 发生之前的某个**前期**气象条件。 使用数值模式，可以预报出未来的**同期**气象条件。

应用举例

事件 A：长江中下游五站平均的当年 6 月降水小于 250mm 的情况

事件 B：长江中下游五站当年 1 月平均降水小于 22mm 的情况

统计 1885-1980 年共 96 年资料统计得：$P(A)=69/96=0.72$　　$P(A/B)=13/14=0.93$

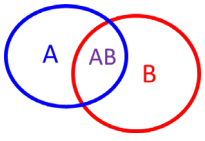
则当 1 月份观测五站平均降水小于 22mm 时，可预报 6 月降水小于 250mm，预报时效 5 个月。

指标判断 条件概率能否作为预报指标的判断必须满足 **2 个经验性条件：**

① $\boldsymbol{P(A/B)\gg P(A)}$ **或** $\boldsymbol{P(A/B)\ll P(A)}$，**差异至少在 0.2 以上**，说明 A 与 B 之间有一定的联系。

② $\boldsymbol{P(A/B)\to 1}$**或** $\boldsymbol{P(A/B)\to 0}$，预报指标有较高的准确率。

3.1.3 事件的独立性



独立事件 如果事件 B 的出现与否不影响事件 A 出现的概率，则称事件 A 对于事件 B 是独立的，满足：$P(A)=P(A/B)$ 或者 $\boldsymbol{P(AB)=P(A)\cdot P(B)}$

① 概率为 0 的事件与任何事件相互独立。

② 若事件A和B相互独立，则$\bar{A}$与B独立，A与$\bar{B}$相互独立，$\bar{A}$与$\bar{B}$也相互独立。

3.3 天气预报指标的统计检验

3.3.1 二项类预报

二项类预报 只预报事件**A出现**或者**不出现$\bar{A}$**，又叫**正反预报**。此类预报，可以有很多预报指标，但**评估其可靠程度**需要了解它们的概率分布。

概率分布 定义符号：在n次独立试验中，事件A出现m次的概率为$P_n(m)$。现在，定义一个事件B，B表示前面A事件发生m次，后面$n-m$发生不发生A事件，既然是独立试验，有：$P(B)=\underbrace{P(A)\dots P(A)}_{m}\times\underbrace{P(\bar{A})\dots P(\bar{A})}_{n-m}$

$P(B)=p^n(1-p)^{n-m}$ ，由于对出现的位置没有限制，那么出现位置的概率是一个组合数，所以有：

$$\boldsymbol{P_n(m)=C_n^mP(B)=C_n^mp^m(1-p)^{n-m}}$$

排列组合的基本知识

布袋中有n个球（每个球都有自己的标号），拿出m个球，（不考虑次序）可能情况的个数是

$$C_n^m=\frac{n\cdot(n-1)\cdot(n-2)\cdot\dots\cdot(n-m+1)}{m!}=\frac{n!}{m!(n-m)!}$$

三个条件

① 每次试验只有 2 个结果：A出现和A不出现。

② 试验条件不变，每次试验A和$\bar{A}$出现的概率是不变的。

③ 试验之间是相互独立的。

应用 计算天气现象出现的概率，特别是**小概率事件**。例如：冰雹、龙卷、超强台风等。

案例：中国体彩大乐透中头奖的概率

彩票规则：35 选 5（前区）加 12 选 2（后区），都选中则获奖。

前区概率：$P(A)=1/35,P(\bar{A})=34/35$，5 个数字都中的概率：$P_5(5)=C_5^5(\frac{1}{35})^5(\frac{34}{35})^0$；

后区概率：$P(A)=1/12,P(\bar{A})=11/12$，2 个数字都中的概率：$P_2(2)=C_2^2(\frac{1}{12})^5(\frac{11}{12})^0$；

最终总的概率为：$P=P_5(5)P_2(2)=1/7{,}563{,}150{,}000$

3.3.2 预报指标的检验

检验的重要性

通过检验，可以认为在设定的标准意义上，**样本的特性可以代表总体的特性**。

3.3.2.1 用二项分布检验天气预报指标

指标检验 预报指标的检验实际上是**反面**来检验该预报指标的**可靠程度**，历史拟合的准确率则从**正面**说明。

检验方法 检验某一条件概率所指示的事件是**属于偶然性**还是**具有规律性**的一种方法，某事件A出现的无条件概率是p，而在条件B时，发现事件A出现的**频率**是$m_{\text{同时发生的次数}}/n_{\text{发生}B\text{的次数}}$，则：

$$\boldsymbol{Q=\sum_{r=m}^{n}C_n^rp^r(1-p)^{n-r}}$$

Q表示在发生B的n次试验中，事件A至少出现m次的概率，**一般小于 0.05 时**我们认为它是一个小概率事件，即**超偶然**。大于 0.05 时，说明这种事很容易发生m次以上，我们认为它偶然出现的概率很大，所以不用条件B做为预报指标。

在利用条件概率进行预报时，我们要求发生的事件不是偶然发生的，就要求其偶然发生的概率非常低，如果在事件B下A发生的偶然发生的概率非常大，说明B指标不能用。

即**当A出现的概率越小，在B中出现的次数越多，Q就越小。**

案例

有云的 10 天中降水 6 天，且降水的无条件概率为 0.2，则 Q 表示任意 10 天（就是观察到B的 10 天，即在B的条件下）中，出现 6 天或以上降水的概率仅仅为 0.637%，这种事件显然不太可能。

如果考虑$p=0.8$（A本身很大），且$P(A/B)=0.95$的情况，就会发现最终结果取决于n的大小：n较大时，说明这 0.15 的提升是显著的，不太可能是因为偶然因素；n较小时，无法确认有效性。

证明与解释

① 假设B条件下，A的发生是偶然的（是无规律的），记为H_1。

② 那么在n次试验中它就容易出现m次及其以上的次数，发生的概率记为Q。

③ 当$Q>0.05$时，A发生m次以上的次数，不是一个小概率事件，H_1成立；说明发生A是偶然的。

④ 当$Q\leq 0.05$时，A发生m次以上的次数，是一个小概率事件，H_1被推翻，说明发生A是非偶然的，是有规律的。

3.3.2.2 小概率事件

小概率事件 随机事件以很小的概率发生，概率很接近于 0（即在大量重复试验中出现的频率非常低）的事件。

小概率标准 一般**多采用 0.01~0.05 两个值**，即事件发生的概率在 0.01 以下或 0.05 以下的事件称为小概率事件。

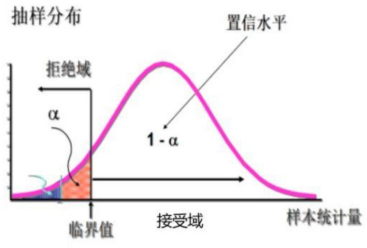
3.3.2.3 小概率事件的应用：假设检验

检验方法 ① **H_0为原假设，H_1为与原假设对立的备择假设**（对立假设）；假设H_0成立；**选定显著性水平α**。

② 构造一个**随机事件A**：当原假设成立时随机事件A以**很小的概率**发生；发现H_0成立是一个小概率事件：从总体中抽取**一个样本**在H_0成立的条件下进行检验，若发现选定的统计量（随机变量）取到此样本代入统计量后的值**是一个小概率事件**，亦即小概率事件在一次试验中发生了，这与小概率原理矛盾。

③ 一般来说，该事件在一次试验中小概率事件不应发生，若发生了，则**否定原假设H_0，接受与其对立的备择假设H_1；反之，接受H_0**

基本思想 首先对**总体的参数或分布形式做出一个假设（原假设）**，然后利用**样本信息**来判断这个假设是否合理。原理是利用小概率事件在一次试验中**几乎是不发生的**来接受或否定假设，是一种有概率性质的反证法。

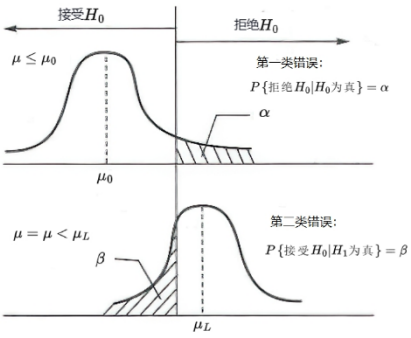


注意两点 ① 这里的**几乎不可能发生**是针对**一次试验**来说的，因为试验次数多了，该事件当然是很可能发生的。

② 当我们运用小概率事件几乎不可能发生的原理进行推断时，**我们也有α%的犯错误的可能**。

两类错误 ① **第一类错误：拒真错误**，即本来原假设是正确的，而根据样本得出的统计量的值落入了拒绝域，拒绝了正确的原假设 **(概率为α)**

② **第二类错误：受伪错误**，即本来原假设是错误的，而根据样本得出的统计量的值落入了接受域，不能拒绝原假设，接受了（确切地说是不拒绝）原本错误的原假设**(概率难以估计)**



两类错误的概率不等，由于第一类错误的概率较小，**一般情况下以拒绝假设的结论为好**。

一般步骤 ① 明确要检验的问题，提出统计假设H_0，与备择假设H_1

② 确定**显著性水平α (0.01, 0.05, 0.1)**（置信水平=1-显著性水平）

③ 在H_0的假设下，针对研究问题，**选取**一个适当的**统计量**；根据观测样本**计算**有关**统计量**；

④ 对给定的 α，根据统计分布，确定事件H_0发生的概率，即确定出临界值，求出H_0的拒绝域；

⑤ 比较统计量计算值与临界值，判断是否显著。

3.4 简单相关系数及检验

状态要素 可以用**条件概率**选择预报因子并且用**二项分布检验**预报因子的可靠程度。

定量数据 主要用**相关系数**选择预报因子或因子集，并用 **t 检验方法**检验其可靠性。

现象间关系 自然界中各现象间存在普遍的关系：

① **确定性关系**：数学上的函数关系　　② **非确定性的关系**：统计上的相关关系（研究出发点）

③ **相关系数**：度量各现象（各要素）间相关程度的量

3.4.1 Pearson 相关系数

背景知识

通常当我们需要用一个数值来表征两个变量（例如：x 和 y）之间的联系。在这种情况下，我们第一个想到的就是使用相关系数。这里的相关系数就是由统计学家卡尔·皮尔逊（Karl Pearson）提出的研究变量之间线性相关程度的统计变量。而变量间的相关程度可以为完全线性相关、线性相关、非线性相关、不相关。

概念

描述两个变量**线性相关程度**的统计量，一般称为**简单相关系数**或**线性相关系数**，用 r 表示，它是两个变量**总体相关系数 ρ** 的估计。因此拿到数据后直接计算相关系数是非常危险的行为，需要先观察分布。两个变量的样本量要求一致，如果不同，必须截去一段。

符号设置

两个变量：$x = x_1, x_2, x_3 \dots x_n \quad y = y_1, y_2, y_3 \dots y_n$ **距平**：$x_i' = x_i - \bar{x} \quad y_i' = y_i - \bar{y}$

协方差和均方差：$s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad s_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

协方差要求变量是同量纲的，相关系数没有要求。标准化后的协方差就是相关系数。

标准化距平：$y_i^* = \frac{y'}{s_y} \quad x_i^* = \frac{x'}{s_x}$

表达式

① **一般表达式**或原始数据表达式：$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \boxed{\frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}x_i y_i - n\bar{x}\cdot\bar{y}}{\sqrt{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^{n}y_i^2 - n\bar{y}^2}}}$

② **距平表达式**：$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}x_i'y_i'}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i')^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i')^2}}$ 平均值为 0

③ **标准化距平**：$r = \frac{1}{n}\sum_{t=1}^{n}x_t^* y_t^*$ 这就是协方差

重要解读

① **相关系数是标准化变量的协方差** ② 其范围在 $-1 \le r \le 1$

③ r 的绝对值越大，表示变量之间关系越密切。

$r > 0$，两变量呈正相关，r 越接近 1，正相关越显著；

$r < 0$，两变量呈负相关，r 越接近-1，负相关越显著；

$r = 0$，两变量不线性相关，相互独立。判断相关系数是否显著，需要通过显著性检验。

$|r| \le 1$的数学证明

从原始定义出发：$r_{xy} = \frac{S_{xy}}{S_x S_y}$，两边平方：$r_{xy}^2 = \frac{s_{xy}^2}{s_x^2 s_y^2}$，由柯西不等式：$\sum_{i=1}^{n}a_i^2 \sum_{i=1}^{n}b_i^2 \ge \left(\sum_{i=1}^{n}a_i b_i\right)^2$，

$\sum_{i=1}^{n}x_i'^2 \sum_{i=1}^{n}y_i'^2 \ge \left(\sum_{i=1}^{n}x_i'y_i'\right)^2 \Longrightarrow s_x^2 s_y^2 \ge s_{xy}^2 \Rightarrow 1 \ge \frac{s_{xy}^2}{s_x^2 s_y^2}$ 因此 $r_{xy}^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} \le 1$ 。

校正

根据统计学中大样本定理，样本量**大于 30** 才有统计意义。

当样本量较小时，计算所得相关系数可能会离总体相关系数甚远。这时，可以用计算**无偏相关系数**加以校正：$r = r\left[1 + \frac{1-r^2}{2(n-4)}\right]$。 然而，绝大多数文献都不会做无偏校正:)。

3.4.2 t 检验

系数检验

正态总体的相关检验实质上是两个变量间或不同时刻间观测数据的**独立性检验**，就是检验总体相关系数$\rho = 0$的假设是否显著。在假设$\rho = 0$成立条件下，样本相关系数r检验的统计量符合**自由度$n-2$的t分布**。所以，可以用t分布检验法来检验。

检验步骤

① 假设$\rho = 0$，计算t值，符合**自由度为$n-2$**的t分布（n为样本量）：

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$

② 给定**信度$\alpha=0.05,0.01$**和**样本相关系数r**，根据自由度相出t_α若

$|t|\geq t_\alpha$，否定$\rho=0$，说明总体是不独立的，总体相关。

$|t|\leq t_\alpha$，接受$\rho=0$，说明总体是独立的，总体非相关。

自由度

统计学上的**自由度 *degree of freedom*** 指样本中可以**自由变动的变量**的个数，当有约束条件时，自由度减少。它包括两方面的内容：**统计量的自由度=样本个数−样本数据受约束条件的个数**，即$df=n-k$（df 自由度，n 样本个数，k 约束条件个数，相关系数中为均值和方差。）

一般来说，自由度等于独立变量减掉其衍生量数，举例来说，变异数的定义是样本减平均值。

例如，一组数据，若平均数一定，则这组数据有$n-1$个数据可以自由变化；若限定了某两数的取值，则自由度为$n-2$。

临界系数

在气象统计预报中，选择因子往往需要计算很多相关系数，逐个如上法检验很麻烦。实际上，在样本量固定情况下，可以计算**统一的判别标准相关系数r_α**，若$r>r_\alpha$，则通过显著性的t检验。

由t_α计算出r_α的计算过程如下：样本容量固定时，通过检验的t值应该至少等于t_α，有$t_\alpha=r_\alpha\sqrt{\frac{n-2}{1-r_\alpha^2}}$，

式中，r_α就是通过检验的相关系数临界值。$r_\alpha=\sqrt{\frac{t_\alpha^2}{n-2+t_\alpha^2}}$　实际应用中，若已知自由度$n-2$和显著性水平，查相关系数表即可。

应用实例

（1）计算得到中国年平均气温与冬季平均气温之间的相关系数为$r=0.48$，且样本量为$n=20$，检验如下：　① 提出原假设：$H_0:\rho=0$，即总体的相关系数为零。

② 计算统计量：$t=\sqrt{20-2}\frac{0.48}{\sqrt{1-0.48^2}}\approx 2.33$

③ 给定$\alpha=0.05$，查询自由度为 18 时的t分布表，由于$t=2.33>t_\alpha=2.10$，故拒绝原假设，认为在$\alpha=0.05$的显著性水平上，中国年平均气温与冬季平均气温之间的相关系数是显著的。

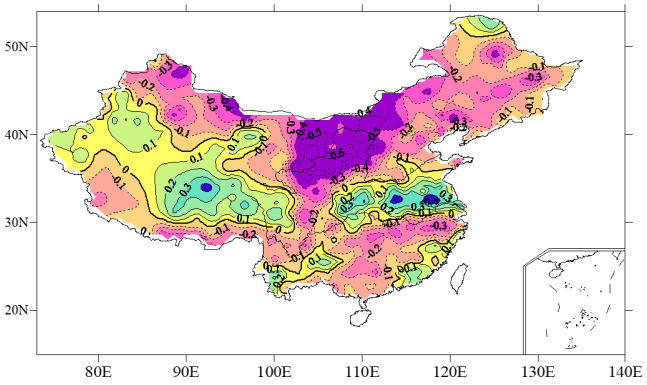
（2）上式数据不变，但使用查相关系数表的办法对r进行检验，自由度 18 对应$\alpha=0.05$时，$r_\alpha=0.44$。由于$r=0.48>r_\alpha$，因此相关系数显著。如果取$\alpha=0.01$，则$r_\alpha=0.56$，不显著。

3.4.3 相关系数在科研中的应用

案例分析

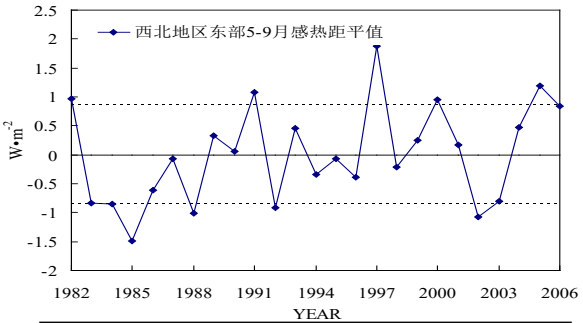
研究我国西北地区热力作用对中国降水的影响：

首先绘制相关系数图（右图），如果想要解释相关系数，可以进一步考虑感热的时间序列（看看信号强的年份是否同样对应显著降水）。信号最强的作为一个样本，最弱的作为一个样本，做合成分析。



西北地区东部夏季感热与全国 634 站同期降水的相关

紫色和深蓝色区域为通过95%置信水平检验的区域，$r_\alpha=0.396$



西北地区东部地面感热	异常偏强年	1982	1991	1997	2000	2005
	异常偏弱年	1984	1985	1988	1992	2002

3.5 复相关系数

单相关矩阵 m个预报因子和预报对象y的单相关矩阵R表示为：$_{m+1}R_{m+1}=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} & r_{1y} \\ r_{21} & r_{22} & \cdots & r_{2m} & r_{2y} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mm} & r_{my} \\ r_{y1} & r_{y2} & \cdots & r_{ym} & r_{yy} \end{bmatrix}$

复相关系数 反映**预报因子集的优劣程度**的数量指标，m个因子与y的复相关系数表示为：$\boldsymbol{R_{y\cdot 12\cdots m}=\sqrt{1-\frac{R^*}{R^*_{yy}}}}$

$R^*=\begin{vmatrix} r_{11} & r_{12} & \cdots & r_{1m} & r_{1y} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mm} & r_{my} \\ r_{y1} & r_{y2} & \cdots & r_{ym} & r_{yy} \end{vmatrix}$　　$R^*_{yy}=(-1)^{2m+2}\begin{vmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mm} \end{vmatrix}$（代数余子式）

特例情况 $m=2$时复相关系数计算公式为：$R_{y\cdot 12}=\sqrt{\frac{r_{1y}^2+r_{2y}^2-2r_{1y}r_{2y}r_{12}}{1-r_{12}^2}}$　复相关系数规定取正值。

m个因子与 y 的复相关系数最大，表示这m个因子线性组合后与 y 的关系最密切。

选取较好的预报因子集要求预报因子之间以及预报因子与预报量之间的相关性如何?

预报因子之间最好相互独立，预报因子集的复相关系数最好大一点。