第二章 气象资料及其表示方法

本章教学重难点

本章将教授气象资料、正态分布统计检验、气候资料的审核和订正的相关内容。

2.1 单个要素的气象资料

补充内容：气象资料的基本概念

用统计方法作气象要素的分析和预报是依据大量的气象观测、模式资料来进行的。

从概论论或统计学的观点来看，某个气象要素及其变化可看成为一个**变量**（或**随机变量**），它的全体在概率论中称为**总体**，而把收集到的该要素的资料称为**样本**。气象统计分析是利用统计学方法对样本进行分析来估计和推测总体的规律性。气象中单个或多个要素可以看为统计学中单个或多个变量。

2.1.1 数据资料

描述 气象资料绝大多数是以数据形式给出的。

表示方法 某气象要素x有n次观测值，向量表达为：$x=(x_1x_2x_3\dots x_n)^T$ 或 $x=(x_t)^T, t=1,2,3,\dots,n$

时间序列 气象要素x在一段时间内的n个观测数据是随时间t变化的序列，称为时间序列。

几何意义 ① n维空间中的一个点 ② 一维空间(单坐标)中的n个点，散点图

2.1.2 统计特征

2.1.2.1 描述平均状况(中心趋势)的统计量

平均值 $\bar{x}=\frac{1}{n}\sum_{t=1}^{n}x_t$ 是要素**总体数学期望**的一个**估计**，反映了该要素的**平均(气候)状况**。气候态

时间平均 包括日平均、月平均、年平均、多年平均值。

空间平均 纬向平均、经向平均、区域平均、半球平均、全球平均等。

大数定律

随机事件的大量重复出现中(30 年以上的平均)，往往呈现几乎必然的规律。

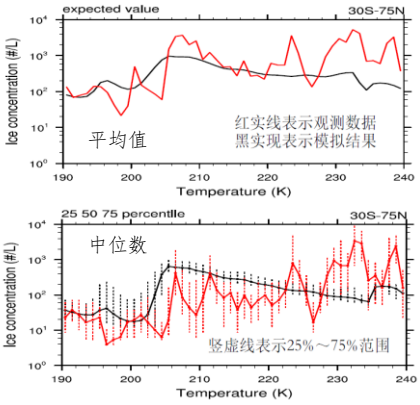
中位数

概念：将变量值按大小顺序排列，处于**中间位置**的那个数就是中位数。

意义：表征变量的中心趋势，反应研究对象的一般水平。

计算：$Position=(n+1)/2$ 若n为偶数，取中间两个数的平均。

优点：不容易受到异常值的干扰，在样本量较小的情况下，这一优点尤为显著。



众数

概念：要素变量值中出现次数最多的数，可以有多个，或者没有

意义：表征研究要素的一般水平，即最容易发生的情况。若变量取值次数较少或取值次数多但无明显集中趋势，计算众数就没有意义。

注意：众数和中位数又称为**位置平均数**，不受到极端变量的影响。

2.1.2.2 描述异常状况(变化幅度)的统计量

距平 $x_t'=(x_t-\bar{x})$ 反映数据偏离平均值的状况，也是通常所说的异常。

距平序列 单个要素样本中每个样本资料点的距平值组成的序列称为**距平序列(距平向量)**。

中心化 **把资料处理为距平**的方法叫做**中心化**。气象上常用距平值代替原样本中的资料值作为研究对象，便于比较且更加直观。平均气温为 36℃或平均气温偏高 2℃。

必要性：① 气象要素在不同周期下平均值不同，为了使他们在同一水平下比较，使用距平值。
② **距平值的平均值为零**，使用方便，直接作为预报值，呈现直观(偏高/偏低)

方差 $s_x^2 = \frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2$ 变量x**减去常数**，方差/均方差**不变**。

标准差 $s_x = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2}$ 标准差与变量值**同量纲**，一般用标准差表示变量取值变化的大小。

意义：描述样本中资料与平均值差异的平均状况，反映变量围绕平均值的平均变化程度(离散程度)。
可以用来评估**波动活跃程度**，方差大的活跃程度大。

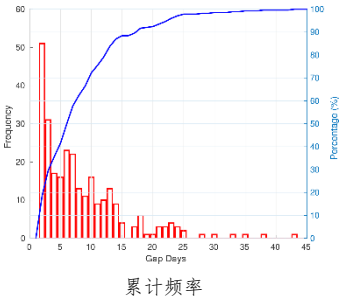
绝对变率 $V_a = \frac{1}{n}\sum_{t=1}^{n}|x_t - \bar{x}|$ **距平绝对值的平均**，说明变量值变化的大小

相对变率 $V_r = \frac{V_a}{\bar{x}}$ **绝对变率**与**平均值**的比值，避免平均值不同的影响

变差系数 $V_p = \frac{s_x}{\bar{x}} = \frac{1}{\bar{x}}\sqrt{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2}$ 表示**变量的相对变化**，是**标准差**与**平均值**的比。

① 绝对变率和标准差的数量级与平均值的量级有关（例如：3 岁的儿童和成年人体重变化）。
② 有些同类型变量，彼此之间平均值差别大（例如不同气候下降水的情况），若要比较它们的变化性用绝对变率和标准差不恰当，应当利用**相对变率**或**变差系数**。

2.1.2.3 其他统计特征



累计频率

频率分布 当两组数据**平均值和均方差相同**(变差系数相同)，但取值有很大区别，为区别两者特征，需要引入**累积频率**。其定义为：变量值小于某上限的次数与总次数之比。

总体 也称为**母体**，是统计分析对象的全体，气象上的总体指**无限总体**。

样本 总体中的一部分，一组气象资料就是无限总体的样本。例如：XXXX 年 XX 月 XX 日的平均气温，一般由 02、08、14、20 时气温的平均值代表。

理解应用 ① 总体的特征是客观存在的，是参数，不是随机变量。
② 样本的特征随所取的样本而变化，与其有关的变量也称为随机变量，如平均值、均方差等。
③ 选取有代表性的样本很重要。
④ 样本量$n \geq 30$ (年)，根据数理统计中的大数定理推断得到。
⑤ 气象上总体指无限总体，一组气象资料就是无限总体的样本。

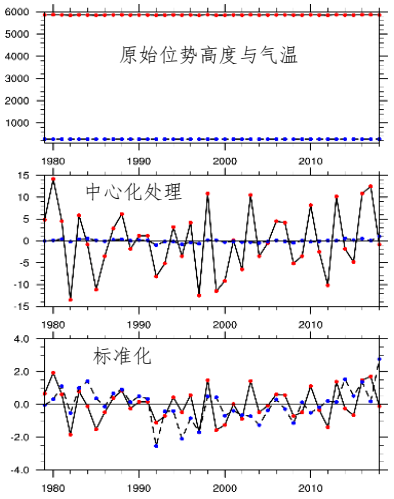
分布函数 **无限总体的累积频率**称为分布函数。

概率密度 $F(x) = P(\xi < x) = \int_{-\infty}^{x} f(x)dx$ $f(x)$称为概率密度函数

正态分布 $f(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ 描述总体特征的μ和σ可以用样本(气象观测数据)平均值和均方差去估计

自然界中不是所有分布都是高斯分布，在研究中可以通过将分布转换为正态分布，以应用相关方法。

2.1.3 数据的标准化



计算公式 $x_t^* = \frac{x_t - \bar{x}}{s_x}$ $t = 1,2,3,\ldots,n$

标准化意义 由于地域差异，各要素**单位不同、平均值和标准差也不同**。为使它们在**同一水平上比较**，采用标准化方法，使它们变成**同一水平的无单位的**变量，即**标准化变量**。

性质 $\overline{x_t^*} = 0$ 标准化变量的平均值为 0 $s_x^* = 1$ 标准化变量的方差为 1。

注意 ① 标准化正态随机变量$P(|x_t^*| > 2.58(1.96)) = 0.01\ (0.05)$
② 标准化变量值的**取值范围在±3之间**，大于 3 的概率仅为 0.0027
③ WMO 旱涝年的确定：距平达到或大于 2 倍标准差（概率不到 5%）
④ 如果资料中出现了极端异常的数值（如标准化后绝对值大于 4），就需要考虑资料的可靠性。

2.1.4 数据的正态化

必要性　各类**统计预报模型**和**统计检验方法**（例如F、t、u、χ^2检验）**要求**数据是符合**正态分布**。

① 月平均气温、气压、多雨地区的月降水量符合正态分布。

② 日降水和少雨地区月降水通常偏态（右偏），旬、候降水则不一定。

处理方法　① **立方根或四次方根**：$x_t'' = \sqrt[3]{x_t}$ 或 $x_t' = \sqrt[4]{x_t}$

② **对数变换**：对原始数据取对数　$x_t' = \ln x_t$

③ **平方根变换**：对离散型变量比较奏效　$x_t' = \sqrt{x_t + 0.5}$

④ **双曲正切转换**：旬降水　$z_t = \text{th}\frac{x_t - \overline{x_t}}{x_t}$

⑤ 化为有序数后的正态化转换（标准化和正态化）

⑥ 角变换（适用于遵从二项分布的变量）：$x_t' = \arcsin\sqrt{x_t}$

状态	暴雨	大雨	中雨	小雨	无雨
频率	3/365=0.0082	13/365=0.0356	41/365=0.1123	100/365=0.274	208/365=0.5699

频率表

2.1.5 状态资料

状态资料　表征气象要素的各种状态，观测结果无法用数据表示。例如降水用“大暴雨”、“大雨”、“中雨”“小雨”和“无”这几个等级来表示；冰雹用“有”、“无”来表示。

统计特征　用**频率表(样本)、分布列(总体)**来描述状态资料的统计特征。列出各个状态出现的频率。对样本而言是频率表，总体而言就是分布列。

2.2 多要素的气象资料

多要素理解　**多个变量、多个格点/站点、多个层次**都可以理解为多要素气象资料。

2.2.1 数据矩阵

表达方法　多个气象要素的样本可以写成矩阵形式。设有***m*个气象要素(维度)**，每个要素有***n*次观测值(样本)**，则任何一个谁可以表示为$\boldsymbol{x_{ij}}$ $(i = 1,2,\ldots m; j = 1,2,\ldots n)$。全部写出并排列为矩阵形式有：

$$X_{mn} = \underbrace{\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}}_{\text{不同时次的观测}} \Big\} \text{不同要素} = (\mathbf{x}_1\mathbf{x}_2\ldots\mathbf{x}_n) = (\mathbf{x}_t)$$

共有n个样本

$\mathbf{x}_t = (x_{1t}x_{2t}\ldots x_{mt})^T$　第t个样本的资料向量，有m个维度

R 型分析　Rows，比较矩阵不同两行，分析不同要素/变量间的关系。

Q 型分析　Queue，比较矩阵不同两列，分析不同时次观测间的关系。

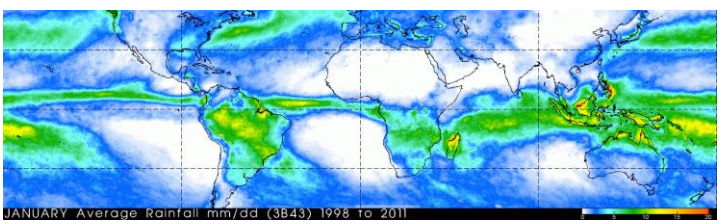
2.2.2 数据的两种空间表示

2.2.2.1 *n*维空间中的*m*个点

R 型分析　每个要素都对应n维空间中的一个点；分析不同要素（变量）之间的关系时用到，如两个变量间的相关关系。　${}_mX_n = (x_1x_2\cdots x_m)^T$　　$x_i = (x_{i1}x_{i2}\ldots x_{in})$　代表n维空间中的一个点

2.2.2.2 *m*维空间中的*n*个点

Q 型分析　每次的观测样本都对应m维空间中的一个点；分析不同时次观测样本之间的关系时用到，如寻找相似个例。　${}_mX_n = (x_1x_2\cdots x_n)$　$x_j = (x_{1j}x_{2j}\ldots x_{mj})^T$　代表m维空间中的一个点



每个格点为一个要素，对格点上历史数据取平均，得到均值矩阵

2.2.3 计算矩阵

2.2.3.1 均值向量

概念 **m个变量的平均值**（n次观测样本的平均）组成的向量。一个变量只有一个均值。

定义 $\bar{\mathbf{x}} = (\bar{x}_1 \bar{x}_2 \cdots \bar{x}_m)^T$　　其中 $\bar{x}_i = \frac{1}{n}\sum_{t=1}^{n} x_{it}$　　$i = 1,2,\cdots,m;\ t = 1,2,\cdots,n$

气象意义 m维空间中的n个点的重心（各部分受到的重力作用集中于一点，这一点就是重心）

举例说明：数据矩阵假设为河北省 10 个城市（10 个变量）近 30 天（30 次观测）的逐日气温，其均值向量表示这 10 个城市每个城市近 30 天的平均气温。

中心化 对矩阵进行中心化处理：$x'_{it} = x_{it} - \bar{x}_i$ 每个值减去对应要素的均值　$_mX'_n = \begin{bmatrix} x'_{11} & x'_{12} & \cdots & x'_{1n} \\ x'_{21} & x'_{22} & \cdots & x'_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x'_{m1} & x'_{m2} & \cdots & x'_{mn} \end{bmatrix}$

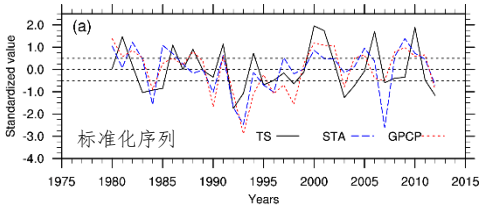
2.2.3.2 协方差和协方差矩阵

协方差 协方差用于**表示两个变量之间的相关关系**

定义 $s_{ij} = \frac{1}{n}\sum_{t=1}^{n}(x_{it} - \bar{x}_i)(x_{jt} - \bar{x}_j) = \frac{1}{n}\sum_{t=1}^{n} x_{it}x_{jt} - \bar{x}_i\bar{x}_j$　　$i,j = 1,2,\cdots m$　**第i行与第j行**

距平形式 $\frac{1}{n}\sum_{t=1}^{n} x'_{it}x'_{jt}$　　向量形式　$= \frac{1}{n}\mathbf{x}_i \cdot \mathbf{x}_j$　　距平向量的点乘/内积

气象意义 反映了**两个变量异常关系**的平均状况，或者两个变量之间的正/负相关关系。两个变量的关系越密切，其协方差的绝对值越大。但是**不同要素之间不好比较，变量需要同单位**。



变量自身的协方差就是方差。

例如：前冬气温正距平（暖）时，后冬气温也出现正距平（暖）；前冬气温负距平（冷）时，后冬气温也出现负距平（冷）。那么，协方差必定为正值，两者间正相关。

案例分析 x_1与x_2的距平符号相同率高，大多具有相同的变化趋势，x_2与x_3的距平符号相反率高，大多具有相反的变化趋势；两组变量均有良好的相关关系。

问题

协方差带单位，不同要素之间不好比较，如果对要素标准化后，再计算协方差，即得相关系数。

协方差矩阵 有m个变量，两两之间共有m^2个协方差（$i = j$时为方差）。这m^2个协方差组成的矩阵称为协方差矩阵：

$$\boldsymbol{S} = (s_{ij}) = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1m} \\ s_{21} & s_{22} & \cdots & s_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ s_{m1} & s_{m2} & \cdots & s_{mm} \end{bmatrix} \quad i,j = 1,2,\cdots m$$

协方差矩阵是**对称矩阵**

矩阵的计算 协方差矩阵可以由矩阵相乘求得。令X^0表示中心化资料组成的矩阵

$$X^0 = \begin{bmatrix} x^0_{11} & x^0_{12} & \cdots & x^0_{1n} \\ x^0_{21} & x^0_{22} & \cdots & x^0_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x^0_{m1} & x^0_{m2} & \cdots & x^0_{mn} \end{bmatrix}$$

其中 $x^0_{ij} = x_{ij} - \bar{x}_i$　　有　$_m\boldsymbol{S}_m = \frac{1}{n}X^0 \cdot (X^0)^T$

矩阵乘法的复习

基本规则：$\mathbf{AB} = \begin{pmatrix} (\mathbf{AB})_{11} & (\mathbf{AB})_{12} & \cdots & (\mathbf{AB})_{1p} \\ (\mathbf{AB})_{21} & (\mathbf{AB})_{22} & \cdots & (\mathbf{AB})_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ (\mathbf{AB})_{n1} & (\mathbf{AB})_{n2} & \cdots & (\mathbf{AB})_{np} \end{pmatrix}$　　其中$(\mathbf{AB})_{ij} = \sum_{k=1}^{m} A_{ik}B_{kj}$

一个例子：$C = AB = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}\begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix} = \begin{pmatrix} 1\times1+2\times2+3\times3 & 1\times4+2\times5+3\times6 \\ 4\times1+5\times2+6\times3 & 4\times4+5\times5+6\times6 \end{pmatrix} = \begin{pmatrix} 14 & 32 \\ 32 & 77 \end{pmatrix}$

离差积　协方差的另一种表示，用离差积表示，构成离差矩阵 $ss_{ij}=\sum_{t=1}^{n}(x_{it}-\bar{x}_i)(x_{jt}-\bar{x}_j)$

2.2.4 整理和利用

多维分布列　针对气象要素的状态资料，统计**多个气象要素**（现象）的**各种情况下的频率**，组成一张**多维频率表**（也称**多维分布列**）。

多维分布列：注意到零点（湿度越高，必然有雨；湿度较低，没有大雨）

降水 / 湿度	无雨	小雨	中雨	大雨	暴雨	合计
14~17(hPa)	0.20	0.04	0.01	0.00	0.00	0.25
18~21	0.16	0.08	0.02	0.02	0.00	0.28
22~25	0.03	0.07	0.12	0.02	0.01	0.25
≥26	0.00	0.01	0.03	0.04	0.14	0.22
合计	0.39	0.20	0.18	0.08	0.15	1.00

整理步骤　一般我们拿到一套空间较大的资料，比如说全球、全国、江苏省的资料，我们需要针对我们的研究范围进行区域性的整理，共有三种方法。

① **代表站方法**：平均相关系数最大的站，该站与各站相关性都很好。

② **区域平均法**：区域平均值要与区域外的格点(站点)值区别大。

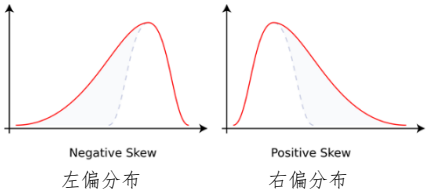
③ **综合指数法**：各站点要素方差差异较大时，同时考虑均值和方差。$K_j=\frac{1}{m}\sum_{i=1}^{m}\left(\frac{x_{it}-\bar{x}_i}{s_i}\right)^2$

2.3 正态分布的统计检验

检验意义　大多数气候诊断方法和预测模型是在气象变量呈**正态分布假定前提**下进行的，所以对气候变量是否呈正态分布形态的检验是十分必要的。正态分布检验不仅可以判断原始变量是否遵从正态分布，还可以检验那些原本不遵从正态分布，但经过数学变换后的变量是否已成为正态分布形式。

偏度系数　$g_1=\sqrt{\frac{1}{6n}}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^3$　其中 s 为均方差，N 为样本量，偏度系数为三阶。

表征**曲线峰点**对**期望值**(平均值)**偏离**的程度(偏态)。是衡量频数分布不对称程度的指标，正态分布时变量值频数以平均数左右完全对称。

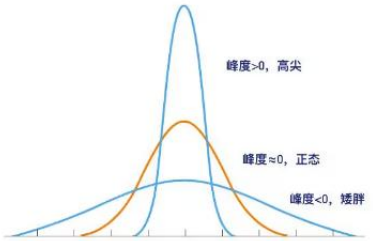


① **偏态系数大于零为右偏分布/正偏分布**，即均值在峰值的右边。

② **小于零为左偏分布/负偏分布**，即均值在峰值的左边，当均值右侧数据较多的时候，均值的左侧必定存在数值较大的离群数据。

峰度系数　$g_2=\sqrt{\frac{n}{24}}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^4-3\right)$　注：不同文献中，计算公式中的系数项（含有样本量n的那一项）可能略有不同。

表征分布形态图形**顶峰的凸平度**(即渐进于横轴的陡度),衡量频数分布的集中程度。**峰态系数小于零为扁平分布，大于零为尖峰分布。**



注意

① $g_1=g_2=0$ 时表示正态分布，但不一定是标准正态。

② 对某一变量作正态性检验，先提出变量遵从正态分布的原假设，计算出g_1,g_2，由此根据标准正态分布表，根据α的水平查出检验结果。

③ 只有峰度和偏度均接受原假设，才可以认为样本来自于正态分布总体。由于正态分布的对称性，查表查的**显著性水平为$\alpha/2$**。

检验依据　已有研究表明，在样本容量很大（近似认为是总体）的情况下，　变量若遵从正态分布。从中任选n个变量（$n\ll$总体的样本量），构成一个样本，样本(样本量n)的偏度系数g_1和峰度系数g_2也都遵从标准化正态分布。因此，可以通过g_1和g_2来判断这个样本是否遵从正态分布。

应用案例

计算得出样本（样本量n）的g_1和g_2：$g_1=0.95$，$g_2=-0.17$；提出变量遵从正态分布的原假设，给定$\alpha=0.05$，则查找0.025的分布函数值。

① 提出原假设H_0：该样本遵从正态分布。

② 给定显著性水平，并查得 $u_\alpha=1.96$（标准化正态分布中，95%的变量的绝对值小于 1.96），由于$g_1=0.96<1.96$，且$|g_2|=0.17<1.96$，因此接受原假设，认为在$\alpha=0.05$显著性水平下，该样本近似遵从正态分布。