第二章 统计量

Robustness：则表明该分析不会受到**数据分布特征**的影响
(μ 数据符合正太则对，数据偏态则错不robust)

Resistance: 则表明它不会"过度"受到数据极值的影响，或者说当数据中的小\较大部分发生变化后，所采用的统计方法计算结果不会发生大的变化。μ 也不resist##

Location

μ

替代平均数μ 更robust/resist 的location统计量：中位数，剪裁平均 $\text{Trimean} = \frac{q_{0.25}+2q_{0.5}+q_{0.75}}{4}$

百分位数: 将数据分布排列，（如中位数，上四分位数，下四分位数等。）

geomean\ harmmean

Spread/Dispersion 变化幅度

距平anomaly

方差（variance） S^2

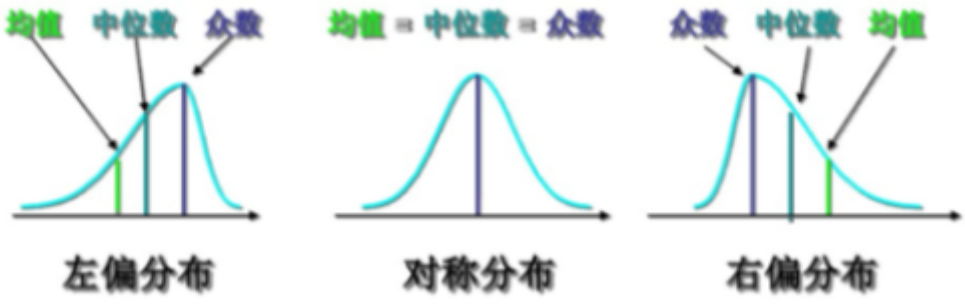
标准差（standard deviation

相比方差 更robust/resist spread统计量：IQR =$q_{0.75} - q_{0.25}$

Symmetry 分布特征统计量

通常用样本的偏态系数来体现数据的分布特征，即相对于中心值的对称性

偏态系数 $\gamma = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^3}{s^3}$



Yule-kendall 指数(更robust)

$$\lambda_{YK} = \frac{(q_{0.75} - q_{0.5}) - (q_{0.5} - q_{0.25})}{IQR} = \frac{(q_{0.25} - 2q_{0.5} + q_{0.75})}{IQR}$$

相关统计量

距平标准化后： 1.无量纲 2.均值0，标准差1

$$z = \frac{x - \bar{x}}{s_x} = \frac{x'}{s_x}$$

相关公式$r_{xy}=\frac{\mathrm{Cov}(x,y)}{s_x s_y}$，上协方差$\mathrm{Cov}(x,y)=\frac{1}{n-1}\sum_{i=1}^{n}\left[(x_i-\bar{x})(y_i-\bar{y})\right]$

Pearson相关则反应了数据对之间**线性**关系的强度
Spearman排序相关很好的体现了数据对之间**单调**关系的强度
自相关 （时间上的+空间上的）
交叉相关

经验分布

柱状图 + 累积频率分布，都是显示哪里数据多的图

符号散点图 在散点上多加了点东西，比如不同颜色表示啥，大小表示啥

相关矩阵

散点图矩阵

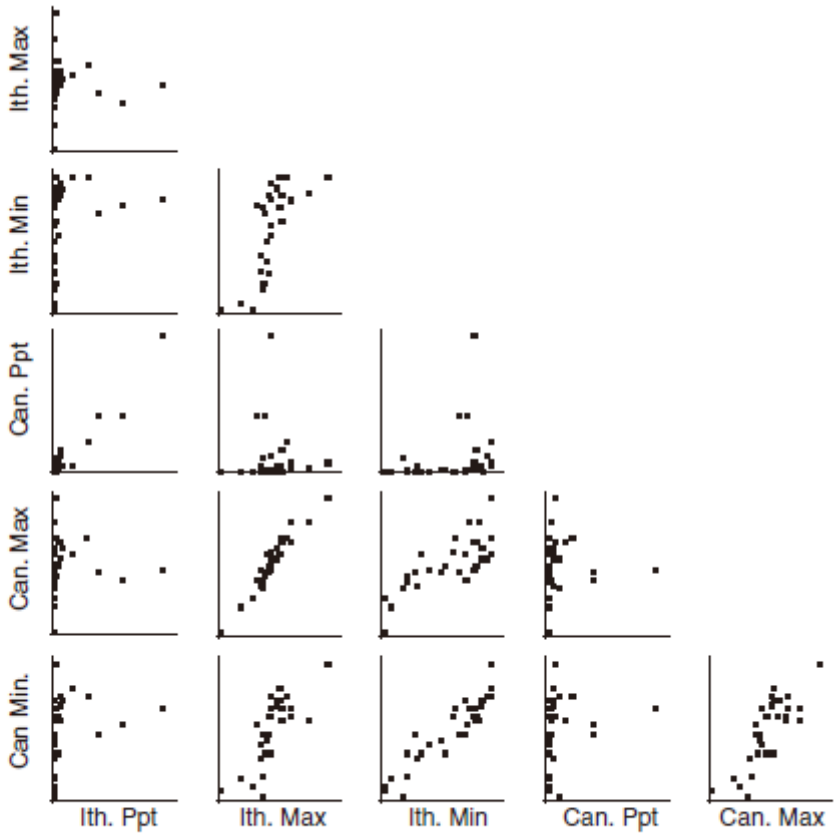
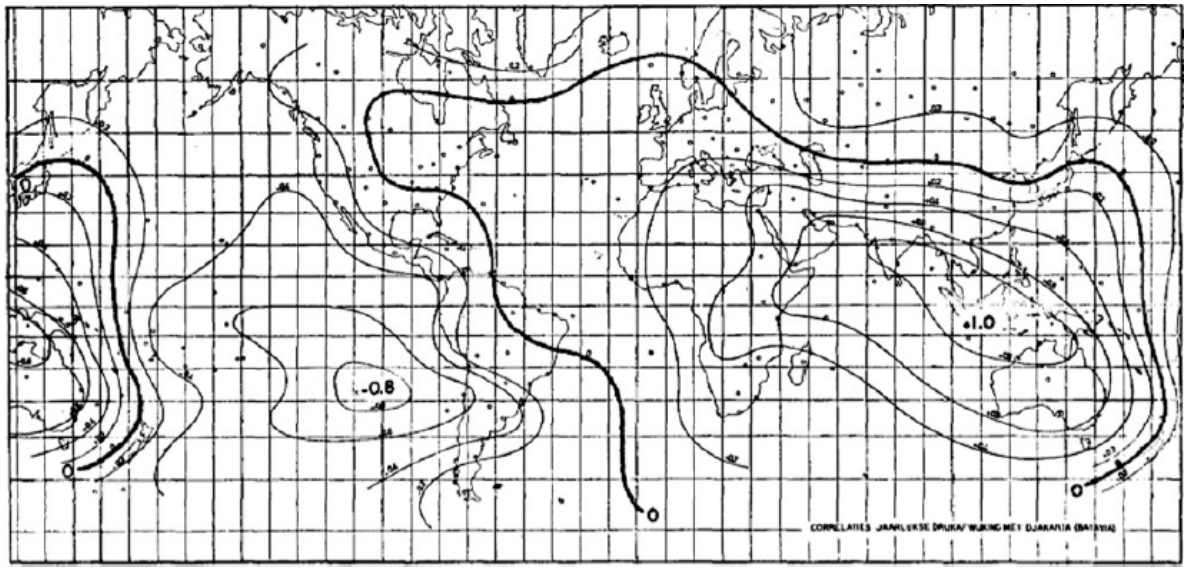


FIGURE 3.27 Scatterplot matrix for the January 1987 data in Table A.1 of Appendix A.

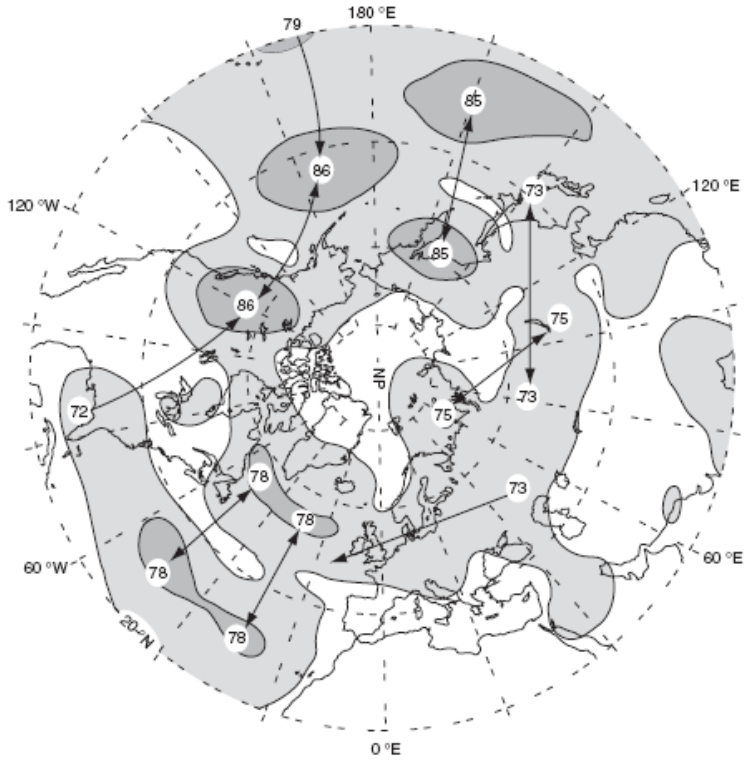
Looking vertically along the column for Ithaca precipitation, or horizontally along the row for Canandaigua precipitation, the eye is drawn to the largest few data values, which appear to line up. Most of the precipitation points correspond to small amounts and therefore hug the opposite axes. Focusing on the plot of Canandaigua versus Ithaca precipitation**, it is apparent that the two locations received most of their precipitation for the month on the same few days**. Also evident is the association of precipitation with milder minimum temperatures that was seen in previous examinations of these same data. The closer relationships between maximum and maximum, or minimum and minimum temperature variables at the two locations—as compared to the maximum versus minimum-temperature relationships at one location—can also be seen clearly.

相关图（相关矩阵升级），一点相关图**One-point correlation map** 空间相关图的相关性在空间上随距离逐渐变弱，但空间上存在遥相关性。



The surprising feature in Figure 3.28 is the region in the eastern tropical pacific, centered on Easter Island, for which the correlations with Djakarta pressure are strongly negative. This negative correlation implies that in years when average pressures at Djakarta (and nearby locations, such as Darwin) are high, pressures in the eastern Pacific are low, and vice versa. This correlation pattern is an expression in the surface pressure data of the El Nin˜o-Southern Oscillation (ENSO) phenomenon, sketched in Example 3.5, and is an example of what has come to be known as a **teleconnection pattern**. In the ENSO warm phase, the center of tropical Pacific convection moves eastward, producing lower than average pressures near Easter Island and higher than average pressures at Djakarta. When the precipitation shifts westward during the cold phase, pressures are low at Djakarta and high at Easter Island.

一点相关图，对相关矩阵的取值，$T_i = |\min_j r_{i,j} \text{ for all } j|$ （P70）



理论分布

优势：

- 压缩性-几个参数就行描述数据，不需要像经验分布对所有数据进行繁杂操作。
- 平滑及内插-数据更连续，不容易受到异常值影响。
- 外推-理论分布可以帮助我们判断气象数据两侧没有数据值可能的发生概率。

离散分布不讲

连续分布

PDF-概率密度函数$\int_x f(x)dx = 1$

CDF-累计分布函数

$F(x) = \Pr\{X \leq x\} = \int_{X \leq x} f(x)dx$
$0 \leq F(x) \leq 1$
$F^{-1}(p) = x(F)$ 直到概率可以反算随机变量

中心极限定理：n十分大，独立同分布数据的算数平均或和服从正态分布（μ，σ_2/n）

大数定律：当试验次数很大时，便可以用事件发生的频率来代替事件的概率。

？？？矩估计

Gaussian $z = \frac{x-\bar{x}}{s}$

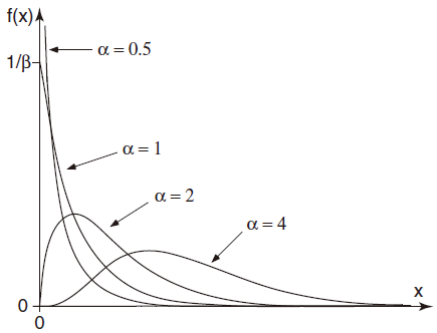
Gamma

α形状参数sharp，β尺度参数scale`

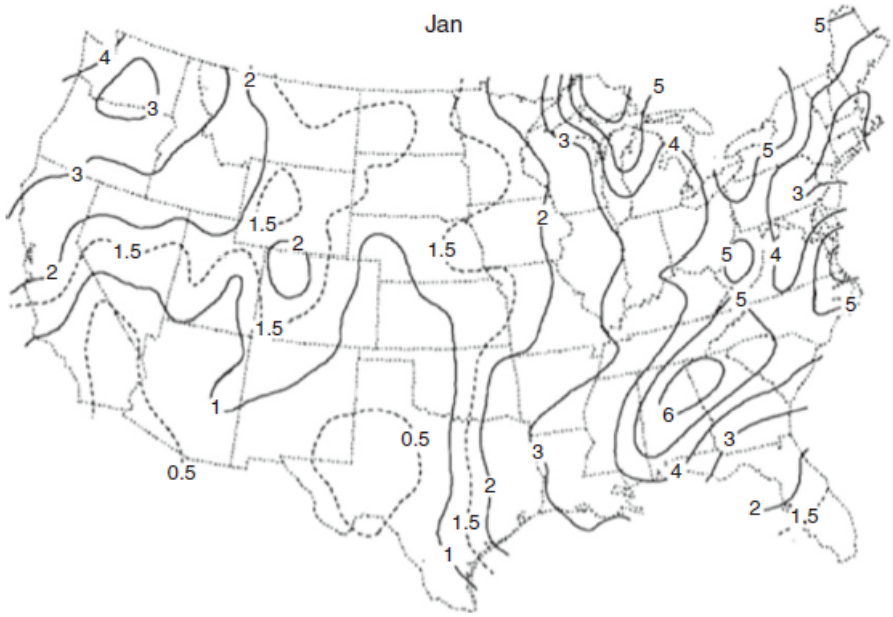
$D = \ln(\bar{x}) - \frac{1}{n}\sum_{i=1}^{n}\ln(x_i)$

$\hat{\alpha} = \frac{0.5000876+0.1648852D-0.0544274D^2}{D}, 0 \leq D \leq 0.5772$
or $\hat{\alpha} = \frac{8.898919+9.059950D+0.9775373D^2}{17.79728D+11.968477D^2+D^3}, 0.5772 \leq \mathrm{D} \leq 17.0$
$\hat{\beta} = \frac{\bar{x}}{\hat{\alpha}}$

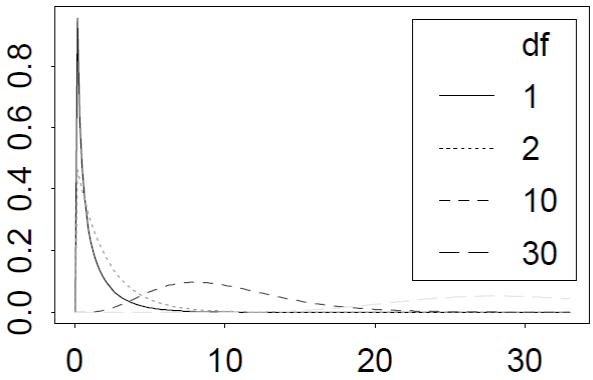
Gamma分布标准化，β=1，α不变，无量纲量：$\xi = \frac{x}{\beta}$

- 

sharp parameter

- 

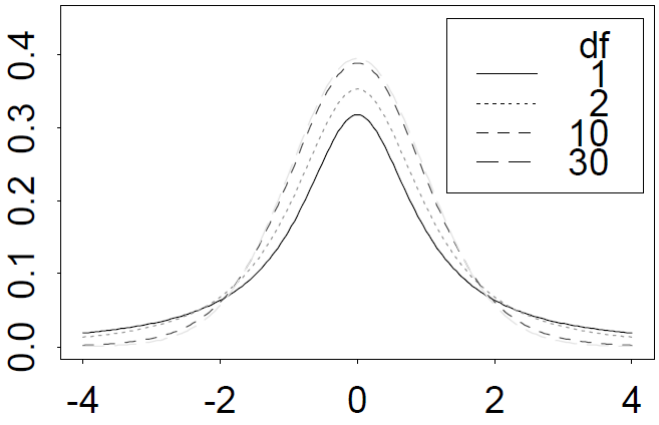
西南部分是明显右偏的降雨分布，而东部更正态分布一些

- α = 1指数分布 雨滴大小
- β=2 χ^2 分布 检验理论分布v.s.实际分布
  - 

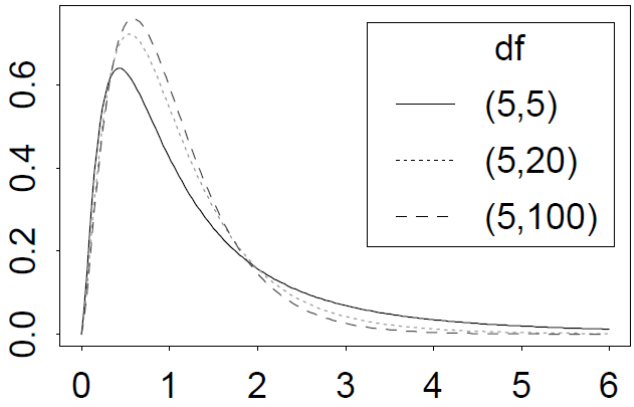
  - 右偏分布，自由度越小偏态性越强，df=30接近高斯分布；

    当df=1，2，pdf峰值位于原点，分布特征依赖于自由度； (df=degree of freedom)

t分布 检验均值



F分布 检验方差

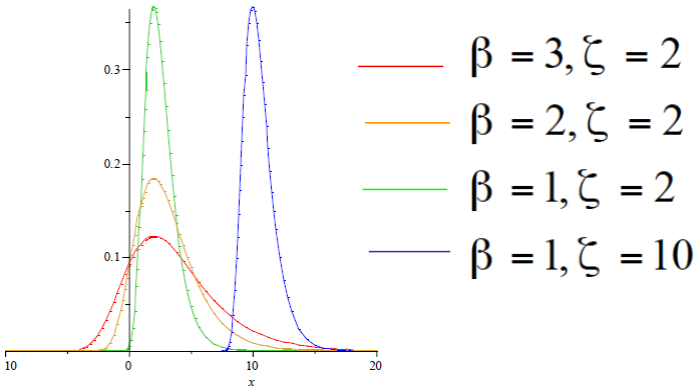


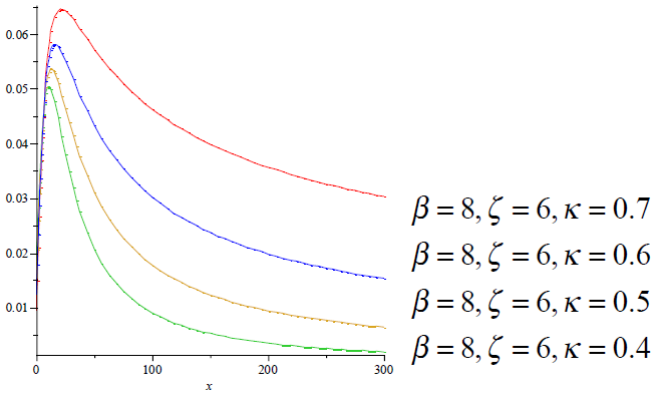
极值分布——block maximum

极端类型定律（Extremal Types Theorem）表明：无论观测数据本身来自于何种固定的单个分布，当独立观测的次数足够多（m的个数），则来自于观测的极端值将遵循某种分布；

1. block maximum: 将数据分段block，每段包含m个数据，选择其中最大的值
2. 方法：**Generalized Extreme Value （GEV）**广义极端值分布

   Location(or shift):ζ
   Scale:β
   Shape:K

   - Gumbel分布 （Fisher-Tippett type I）
     - K趋近0
     - 

   - Frechet（Fisher-Tippett type II）
     - K>0
     - 特殊 heavy tail现象
     - 

   - Weibull（Fisher-Tippett type III）
     - K<0 风速
     - α = 1，等同于指数分布
     - α = 3.6， 类似于高斯分布

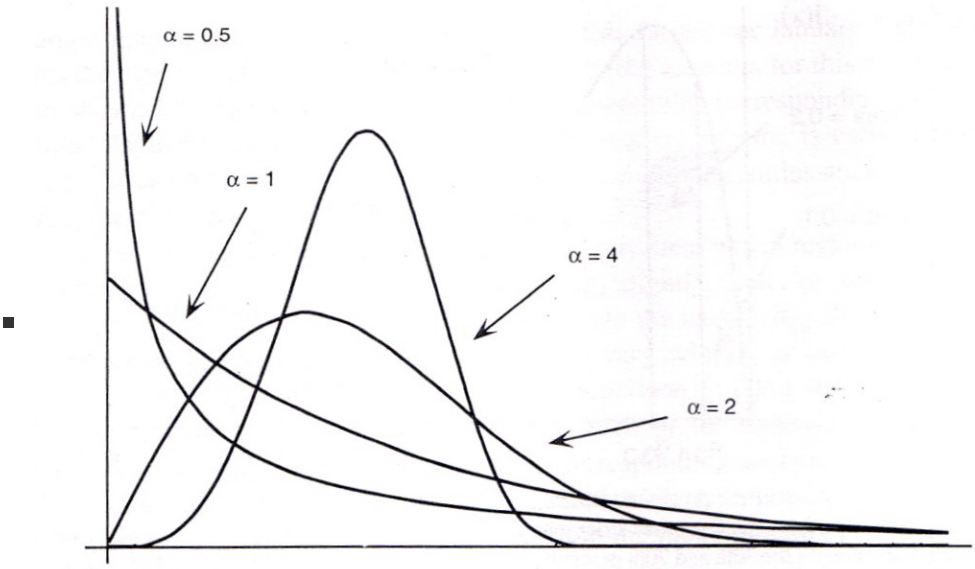
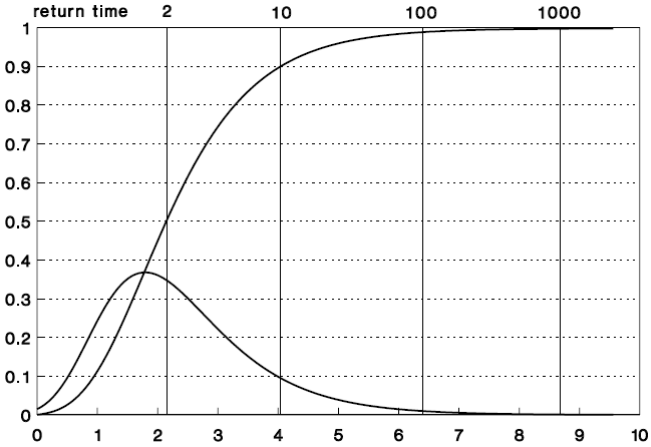


Fig. 4.10 Weibull distribution probability density functions for four values of the shape parameter, α.

3. **使用注意事项**
    - 数据独立，且来自于同一分布，以及观测数据m足够大，但通常较难满足；
    - 即使不满足上述条件，GEV也可以用，但不能保证拟合的效果；
    - block maximum方法的缺点：可能造成大量数据信息丢失；

Return Value

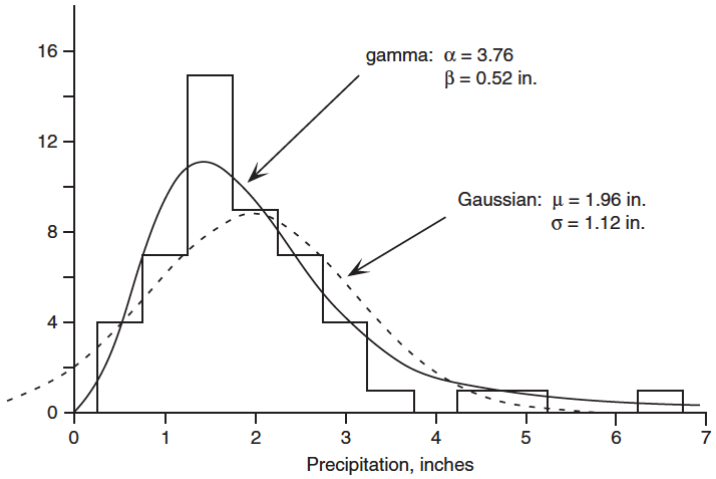
某次事件再次出现的时间长度（如平均或超过5、10、50年再次出现，等等），估计其在极值分布中对应的阈值，即关注极值分布中**上百分位数**；



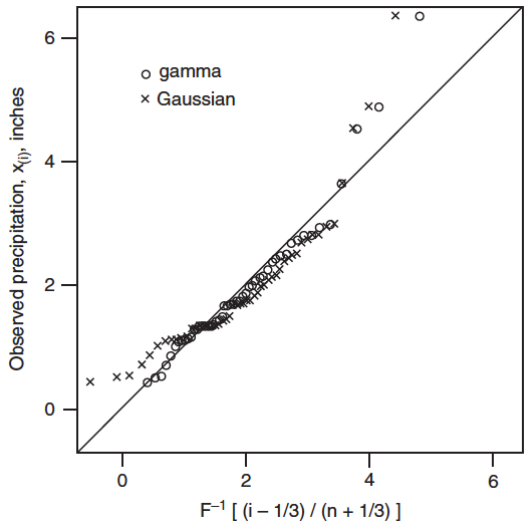
某一Gumbel极值分布（年最大值数据）的PDF和CDF，垂直线代表2，10，100以及1000年的return value；此处2年的return value对应的F(x)=50%, 10年的return value对应的F(x)=90%(看CDF)

QQ-Plot

Quantile–quantile (Q–Q) plots compare empirical (data) and fitted CDFs in terms of the dimensional values of the variable (the empirical quantiles).



Apparently, the gamma distribution provides a reasonable representation of the data. The Gaussian distribution underrepresents the right tail and implies nonzero probability for negative precipitation.



Quantile–quantile plots for gamma (○) and Gaussian (x) fits to the 1933–1982 Ithaca January precipitation in Table A2. Observed precipitation amounts are on the vertical, and amounts inferred from the fitted distributions using the Tukey plotting position are on the horizontal. Diagonal line indicates 1:1 correspondence.

第三章 参数检验

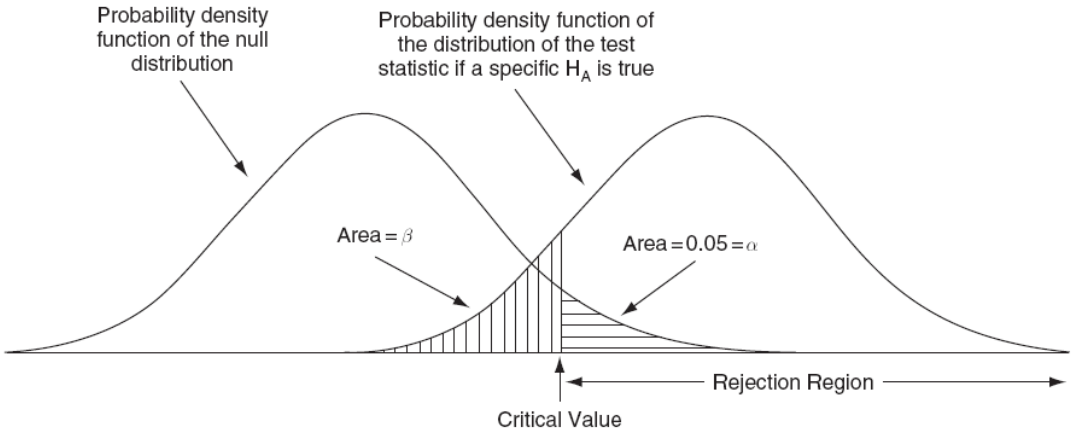
概念

检验水平(test level/level)——α： 零分布中足够说明不可能发生的区域。0.01，0.05，0.1。

P值：样本满足零分布的前提下，样本计算得到的检验统计量的具体概率值。（零分布是，零假设里的分布）

二类错误

第一类错误：弃真，第二：纳伪 一般指控制α第一类错误的范围。



均值检验

单个值

u检验，$u=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$，方差已知情况

t检验，方差未知；优点：小样本

$$t=\frac{\bar{x}-\mu_0}{[\mathrm{Var}(\bar{x})]^{1/2}} \quad \mathrm{Var}[\bar{x}]=\frac{s^2}{n}$$

成对数据检验

正态分布检验，方差已知情况

t分布检验：方差未知（小样本）

- 方差不等

$$\begin{aligned}\mathrm{t}&=\frac{\bar{x}_1-\bar{x}_2}{\left[s_1^2/n_1+s_2^2/n_2\right]^{1/2}}\\ v&=\min\left(n_1,n_2\right)-1\end{aligned}$$

- 方差等

$$\begin{aligned}\mathrm{t}&=\frac{\bar{x}_1-\bar{x}_2}{\left[s_p^2/n_1+s_p^2/n_2\right]^{1/2}}\\ s_p^2&=\frac{df_1 s_1^2+df_2 s_2^2}{df_1+df_2}\\ df&=df_1+df_2=n_1+n_2-2\end{aligned}$$

正态分布：大样本（避免考虑样本方差齐性）

$$\mathrm{t}=\mathrm{Z}=\frac{\left(\bar{x}_1-\bar{x}_2\right)-\left(\mu_1-\mu_2\right)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

同时观察数据检验问题（地点相关性）

大气科学中同时观测两地的数据，直接检验，过高估计方差，统计值会降低，所以会容易认为均值相等。（应该被拒绝，却接受了）
Solution：双值相减变单值，即差，然后检验。

时间非独立数据检验（自相关性）

气象中的持续性（自相关）使得数据时间平均的方差比独立数据大，因此在使用前面所给出的方法分析通常会**“低估”**统计检验分布的方差部分，从而增大了统计检验的值，因此增大了平均值差异通过显著性检验的可能性。

Solution: 要先进行滤波。要考虑方差膨胀。

方差检验

单样本

χ^2检验

μ不知：$\chi^2=\frac{(n-1)s^2}{\sigma^2}$　自由度　$v=n-1$

μ已知：$\chi^2=\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2$ 自由度，$v=n$

双样本

F检验

μ_1,μ_2不知:$F=\left(\frac{n_1}{n_1-1}s_1^2\right)/\left(\frac{n_2}{n_2-1}s_2^2\right)$

相关系数的检验

t检验

原假设为0

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

自由度$v=n-2$

拟合优度 Goodness of Fit Tests

χ^2检验

- 比较经验和理论的**PDF**或离散分布函数
- 事先需要将数据分为离散的数据组，因此更合适用于分析离散随机变量；
- 对于连续变量，数据存在四舍五入分组的情况，可能会造成严重的信息损失的现象

  方法：

  分组原则：每一组中得到的期望数据个数不能太小（5个以上）；
  　　每组的概率或者范围不一定相等；

  右侧检验

$$\begin{aligned}\chi^2&=\sum_{\text{classes}}\frac{(\#\text{ Observed }-\#\text{ Expected })^2}{\#\text{ Expected}}\\&=\sum_{\text{classes}}\frac{(\#\text{ Observed }-n\Pr\{\text{ data in class }\})^2}{n\Pr\{\text{ data in class }\}}\end{aligned}$$

  检验自由度为

$$v = (\#\text{ of classes - }\#\text{ of parameters fit } - 1)$$

例子

例如，分别用Gamma和Gaussian分布拟合1933-1982年一月Ithaca的降水；
Gamma分布的参数：α = 3.76， β = 0.52 in
Gaussian分布参数： μ = 1.96in， σ = 1.12in

TABLE 5.1 The χ^2 goodness-of-fit test applied to gamma and Gaussian distributions for the 1933–1982 Ithaca January precipitation data. Expected numbers of occurrences in each bin are obtained by multiplying the respective probabilities by $n = 50$.

Class	$<1''$	1 — $1.5''$	1.5 — $2''$	2 — $2.5''$	2.5 — $3''$	$\geq 3''$
Observed #	5	16	10	7	7	5
Gamma:						
Probability	0.161	0.215	0.210	0.161	0.108	0.145
Expected #	8.05	10.75	10.50	8.05	5.40	7.25
Gaussian:						
Probability	0.195	0.146	0.173	0.173	0.132	0.176
Expected #	9.75	7.30	8.65	8.90	6.60	8.80

自由度 $6 - 2 - 1 = 3$,显著性0.1情况下。
$\chi_1^2 = 5.05$, $\chi_2^2 = 14.96$

Gamma: $\chi^2_{0.1}(3) = 6.251$不拒绝
Gaussian: $\chi^2_{0.01}(3) = 11.345$ 拒绝， $\chi^2_{0.001}(3) = 16.345$ 不拒绝

K-S检验

- 比较的是经验和理论分布的**CDF**
- 对于连续分布，K-S检验通常比χ^2检验更有用；
- 单侧检验；原假设该理论分布可行
- 局限性：
  - 分布参数不由样本估计得到（否则拟合的很好）
  - 检验重点——检验临界值选择（通常会造成本应拒绝零假设的检验接受了零假设。）
  - 临界值选择：（旧的，不看了）

$$C_\alpha = \frac{K_\alpha}{\sqrt{n} + 0.12 + 0.11/\sqrt{n}}$$
$$K_\alpha = 1.224, 1.358, \text{ and } 1.628, \text{ for } \alpha = 0.10, 0.05, \text{ and } 0.01$$

  - Lilliefors的K-S检验统计量 $D_n = \max|F_n(x) - F(x)|$
    - 经验累积概率，$F_n\left(x_{(i)}\right) = i/n$； F(x) 理论累积分布函数
    - **临界值依赖于所选择的分布**
      - Gamma分布：检验临界值依赖于样本容量n和参数α，
      - Gaussian分布：α = 无穷

例子

例：分别用Gamma和Gaussian分布拟合1933-1982年一月Ithaca的降水；

- 比较月降水量的经验累积概率和理论CDF；
  寻找差异——Gamma和Gaussian分布检验的最大差异点为同一点：
- Gamma $D_n = 0.068$
  Gaussian $D_n = 0.131$

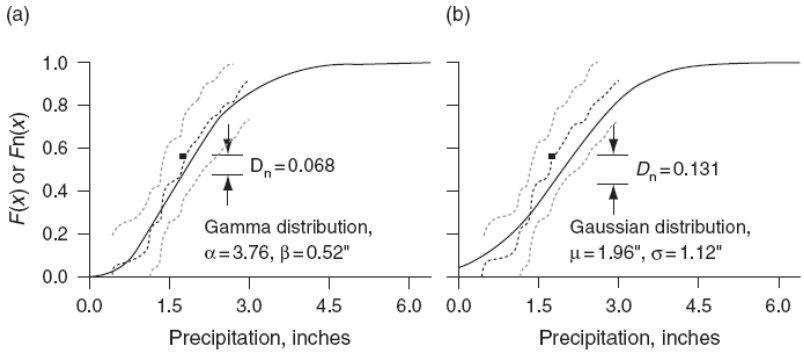


FIGURE 5.5 Illustration of the Kolmogorov-Smirnov D_n statistic as applied to the 1933–1982 Ithaca January precipitation data, fitted to a gamma distribution (a) and a Gaussian distribution (b). Solid curves indicate theoretical cumulative distribution functions, and black dots show the corresponding empirical estimates. The maximum difference between the empirical and theoretical CDFs occurs for the highlighted square point, and is substantially greater for the Gaussian distribution. Grey dots show limits of the 95% confidence interval for the true CDF from which the data were drawn (Equation 5.16).

Filliben Q-Q csorrelation Test

方法：把Gaussian分布数据转换成百分位那种F，（$p_i = \frac{i-1/3}{n+1/3}$，Tukey）
计算相关F和原始数据的相关系数，r和下面的值对比（Critical Value）

TABLE 5.3 Critical values for the Filliben (1975) test for Gaussian distribution, based on the Q-Q plot correlation. H_0 is rejected if the correlation is smaller than the appropriate critical value.

n	0.5% level	1% level	5% level	10% level
10	.860	.876	.917	.934
20	.912	.925	.950	.960
30	.938	.947	.964	.970
40	.949	.958	.972	.977
50	.959	.965	.977	.981
60	.965	.970	.980	.983
70	.969	.974	.982	.985
80	.973	.976	.984	.987
90	.976	.978	.985	.988
100	.9787	.9812	.9870	.9893
200	.9888	.9902	.9930	.9942
300	.9924	.9935	.9952	.9960
500	.9954	.9958	.9970	.9975
1000	.9973	.9976	.9982	.9985

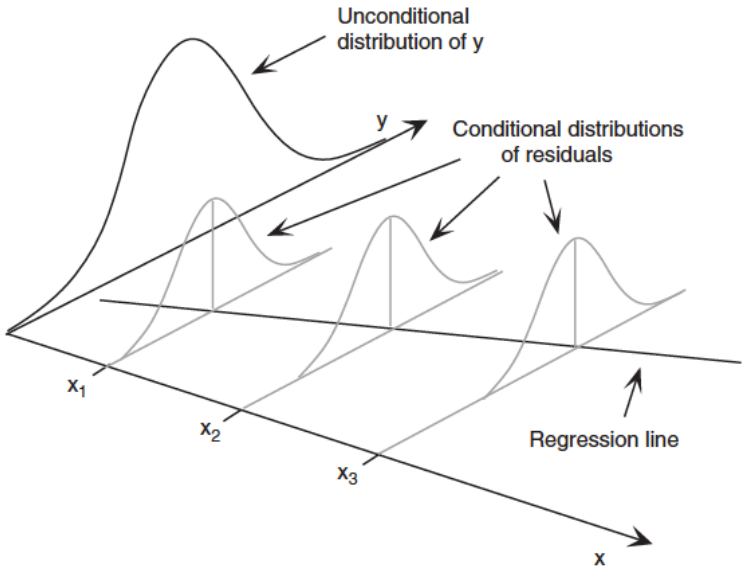
第四章

一元线性回归

残差特点，方差分析表，条件

残差分析

线性回归的残差应满足以下几个条件：
1) **是独立的随机变量**；
2) 数学期望为0；
3) 方差为常数；
4) 满足正态分布；



方差分析

样本总平方和 $SST=\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2=\sum_{i=1}^{n}y_i^2-n\bar{y}^2$
样本回归平方和$SSR=\sum_{i=1}^{n}\left(\hat{y}_i\left(x_i\right)-\bar{y}\right)^2$
残差平方和$SSE=\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2$

$$SST=SSR+SSE$$

R^2=SSR/SST （=r_{xy}^2）被称为解释方差:反应了两个变量之间的线性关系密切程度程度。

TABLE 7.1 Generic analysis of variance (ANOVA) table for simple linear regression. The column headings df, SS, and MS stand for degrees of freedom, sum of squares, and mean square, respectively. Regression df = 1 is particular to simple linear regression (i.e., a single predictor *x*). Parenthetical references are to equation numbers in the text.

Source	df	SS	MS	F
Total	$n-1$	SST (7.12)		
Regression	1	SSR (7.13)	MSR = SSR / 1	(F = MSR/MSE)
Residual	$n-2$	SSE (7.14)	MSE = s_e^2	

平均回归平方和$MSR=SSR/1$，自由度1
残差方差$MSE=(SST-SSR)/(n-2)$,自由度 n-2

显著性检验，$R^2=\frac{SSR}{SST}=r_{xy}^2$，它不决定拟合好坏

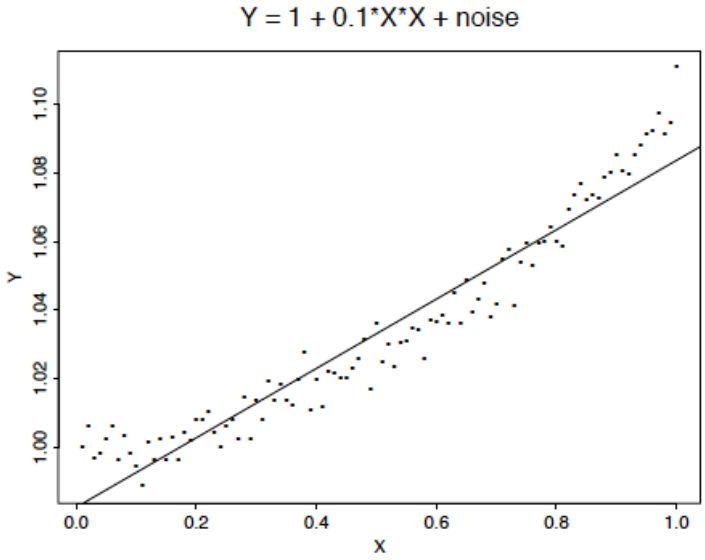


Figure 8.6: *This diagram illustrates the least squares fit of a straight line to a sample of 100 observations generated from the model* $\mathbf{Y} = 1 + 0.1\mathbf{x}^2 + \mathbf{E}$ *where* $\mathbf{E} \sim \mathcal{N}(0, 0.005^2)$. *Even though* $R^2 = 0.92$, *the model fits the data poorly.*

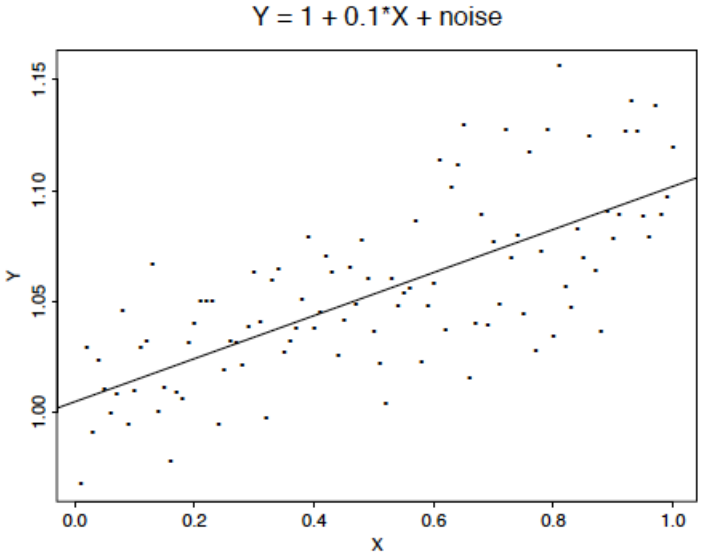


Figure 8.7: *This diagram illustrates the least squares fit of a straight line to a sample of 100 observations generated from the model* $\mathbf{Y} = 1 + 0.1\mathbf{x} + \mathbf{E}$ *where* $\mathbf{E} \sim \mathcal{N}(0, 0.025^2)$. *Even though* $R^2 = 51\%$, *the model fits the data well.*

回归方程的F检验

在原假设**回归系数为0**的条件下（即不存在线性回归关系）统计量

$$F = \frac{SSR/1}{SSE/(n-2)} = \frac{MSR}{MSE}$$

服从分子自由度为1，分母自由度为(n-2)的F分布

例如： 给定置信度95%，查F分布表，分子自由度为1，分母自由度为(20-2)的$F_{\alpha=0.05}$=4.41，F = 20.18 则有表明二者的线性回归关系是显著的。

第五章

气候时间序列：

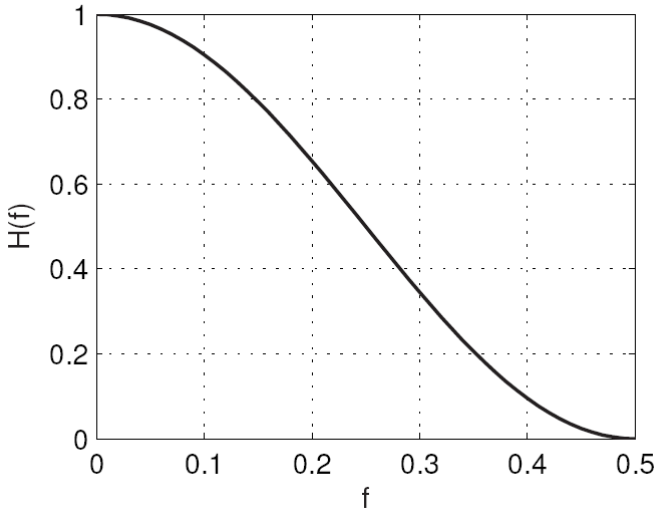
- 随时间变化的一列气候数据构成了一个气候时间序列；气候时间序列的趋势是指气候要素大体的变化情况，即描述很长时间尺度的演变过程；

一般特征：

- 数据的取值随时间变化；
- 每一时刻取值具有随机性；
- 前后时刻数据之间具有相关性和持续性；
- 序列整体上有上升或下降趋势，并呈现周期性振荡；
- 某一个时刻数据取值可能出现转折或突变；

提取方法：三次样条法，滑动平均5点二次，7点二次，9点二次。

3点加权滑动平均（1-2-1）的频率响应函数



对于无限大的周期 f→0（趋势），频率响应H(f) → 1 ，表示在过滤后无任何削弱；

高频部分，f=0.5，即周期为2的分量已完全消除；

检验方法：

- Z检验，0.05显著性，在拒绝域内会就显著
- Mann-Kendall Trend Test,要求无自相关（红噪声），白噪声去除（数据不够+有显著lag1自相关）
- SR test
- Sen's slope 斜率判断是否有单调趋势
  固有周期：有规律的周期变化，年循环、季节变化
  准周期：会偶尔出现的循环，周期无法确定，比如厄尔尼诺。

气候突变检验

滑动t检验:把两端子序列的μ，看作两个总体μ，来t检验
Yamamoto method:信噪比来t检验
sequential Mann-Kendall test(SQ-MK) 正秩、反秩的交点就是突变点。优点是不需要考虑子段长度；但存在多个跃变点，不合适
Pettitt 构造秩序列，寻找最大值，带公式，也不适合突变点较多的情况

第七章

对已经存在的时间序列用公式表达，来得到其振荡周期，振幅等数据。这点和时间序列提取平滑不同。

弱平稳时间序列:
是指其中随机变量的时间序列,它的前期演变过程的统计相关规律在未来的一段时间内是不变的:

- 数学期望值与方差是不变的;
- 它的相关函数只与时间间隔有关而与时间无关（弱平稳或协方差平稳）;

  平稳化处理方法：1.去除年循环 2.数据分级：当序列足够短的时候就可以将它近似看作平稳的。

时间序列分析方法

- 时间域分析方法（离散：Markov；连续：自回归）
- 频率域分析方法：谐波分析、谱分析

1. 谐波分析（Harmonic analysis）：

谐波分析是将一系列sine和cosine函数叠加在一起来表征原始数据的振荡或波动（midlatitude降雨数据=diurnal cycle+annual timescale)

$\alpha = \frac{t}{n}2\pi$
$y_t = \bar{y} + C_1\cos\left(\frac{2\pi t}{n}\right)$ $\bar{y}$上下移动距离，C_1压缩拉伸
$y_t = \bar{y} + C_1\cos\left(\frac{2\pi t}{n} - \phi_1\right)$ ϕ_1相位角

谐波与多元线性回归

$$\cos(\alpha - \phi_1) = \cos(\phi_1)\cos(\alpha) + \sin(\phi_1)\sin(\alpha)$$
$$C_1\cos\left(\frac{2\pi t}{n} - \phi_1\right) = C_1\cos(\phi_1)\cos\left(\frac{2\pi t}{n}\right) + C_1\sin(\phi_1)\sin\left(\frac{2\pi t}{n}\right)$$
$$= A_1\cos\left(\frac{2\pi t}{n}\right) + B_1\sin\left(\frac{2\pi t}{n}\right)$$
$$A_1 = C_1\cos(\phi_1)$$
$$B_1 = C_1\sin(\phi_1)$$
$$x_1 = \cos\left(\frac{2\pi t}{n}\right), x_2 = \sin\left(\frac{2\pi t}{n}\right)$$
$$A_1 = b_1, \quad B_1 = b_2$$

用最小二乘法推出公式到A_1 ,B_1, 用表算出带入得到值

$$A_1 = \frac{2}{n}\sum_{t=1}^{n} y_t\cos\left(\frac{2\pi t}{n}\right)$$
$$B_1 = \frac{2}{n}\sum_{t=1}^{n} y_t\sin\left(\frac{2\pi t}{n}\right)$$

过度拟合：多元回归中，当拟合线通过所有数据点时，复相关关系为100%.若谐波方中包含n/2个谐波时，也称过度拟合。~~n/2能把所有点包含进去。~~分析中是可以过度拟合的，预报不可以

2. 谱分析(spectral analysis)

2.1. **原因1**：谐波函数彼此独立的特性，来自于sine和cosine函数彼此正交性;

单个谐波方差贡献：由于每个谐波之间彼此独立，因此它们对方程的方差贡献并不随谐波方程的变化而变化，如对第k个谐波而言，其方差贡献为:（**功率谱纵坐标**）

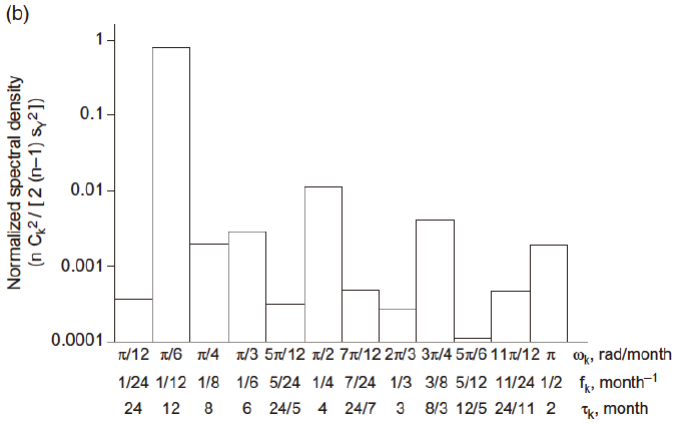
$$R_k^2 = \frac{\frac{n}{2}C_k^2}{(n-1)s_y^2}$$

当有n/2个谐波，R^2=1, S_y^2原序列样本方差。注意，它常会被以**对数**方式表示。

2.2. **周期图/Fourier线谱**（功率谱power）

- 定义： 时间频率因Fourier转换到频率域上的图像，可以图像分析，这个图像叫周期图或Fourier线谱。
- 离散功率谱（表达形式，纵坐标）
  - 振幅 $C_k^2 = \omega_k$，横坐标还可以是频率$f_k = \frac{k}{n} = \frac{\omega_k}{2\pi}$， （1/n,1/2），周期$T_k = \frac{n}{k} = \frac{2\pi}{\omega_k} = \frac{1}{f_k}$，波数k
  - 标准谱密度，R_k^2
- 意义：

  – 提供了不同频率谐波对原数据的贡献；
  – 但并没有提供位相角的信息，即没有提供不同频率谐波随时间的变化信息， 从而无法重构时间序列



Nyquist频率 $k = \frac{n}{2}$时$\omega_2 = \pi$, 最快频率1波下。

Aliasing假名问题： 可能序列存在的重要的物理过程频率比Nyquist频率更高频，但这些短周期振荡无法由直接分辨出来，则它们的作用会体现在较长周期（ ω_1和ω_2之间）中；也就是高频信号被冒名顶替了的现象。

Aliasing形成原因:

• 实际时间序列可能存在比Nyquist频率更高的快变化部分；
• 数据取样间隔过大，则不能体现这种快变化； 而这种比Nyquist频率高的频率并不会因此从数据真实变化中消失；
• 则这些高频的作用便虚假的体现在可分辨的频率中

如何避免Aliasing:

—但选定数据，则没有办法去除“Aliasing”现象，但:
– 通过提高数据分辨率的方法尽可能的避免该现象；
– 或者根据已知的物理过程，来确定资料/样本的变化率，从而达到去除假名现象；
– 对于探索性的研究，即不知道物理本质的问题，是没有办法去除假名现象的，可期望接近Nyquist频率的功率谱能量接近0，则可能说明高频部分的能量很小；

白噪声：表示不含有任何规律性波动的随机过程。由强度相同的各种频率振荡共同组成的随机序列。

红噪声：泛指一种含极长波长的红外光所组成。随频率增加噪声能量单调递减。

第八章

主成分分析PCA (principal component analysis)

对比回归分析：采用多个因子（方差分析），但因子间可能存在相关造成多余信息，于是预报效果会差。

EOF分析特点与优势：

- 则是利用最少的EOF因子解释数据集最大程度的变化（方差）；
  从客观的角度探寻数据集的变化结构；
  分析变量间的关系；
- EOF的结构基于数学方法获得，不一定对应某特定物理含义；
  因此对其结果的描述要求基于物理事实或直觉；

经验正交函数分解（EOF）

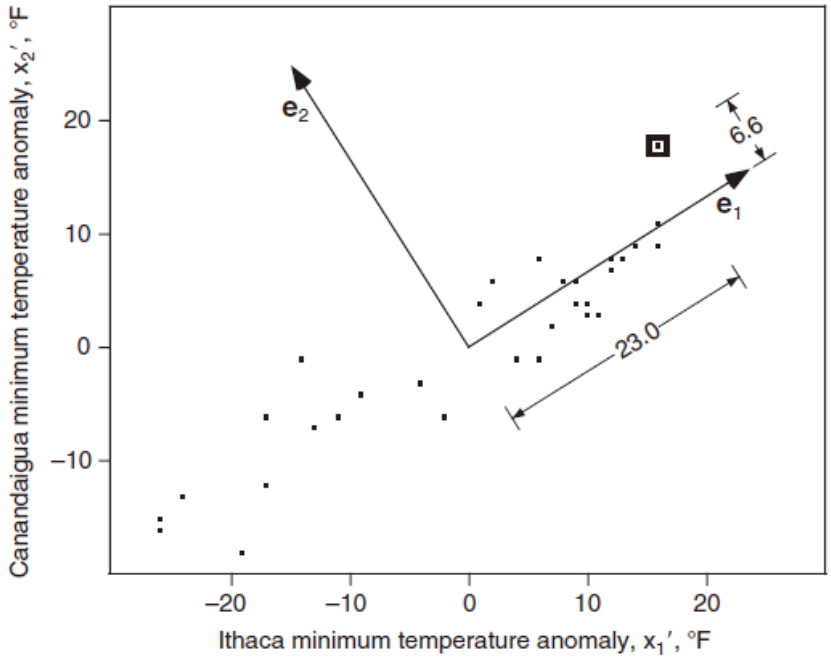


FIGURE 12.1 Scatterplot of January 1987 Ithaca and Canandaigua minimum temperatures (converted to anomalies, or centered), illustrating the geometry of PCA in two dimensions. The eigenvectors $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ of the covariance matrix $[S]$ for these two variables, as computed in Example 10.3, have been plotted with lengths exaggerated for clarity. The data stretch out in the direction of $\boldsymbol{e}_1$ to the extent that 96.8% of the joint variance of these two variables occurs along this axis. The coordinates u_1 and u_2, corresponding to the data point $\boldsymbol{x}'^T$ [16.0, 17.8], recorded on January 15 and indicated by the large square symbol, are shown by lengths in the directions of the new coordinate system defined by the eigenvectors. That is, the vector $\boldsymbol{u}^T = [23.0, 6.6]$ locates the same point as $\boldsymbol{x}'^T = [16.0, 17.8]$.

可以看到Ithaca变化幅度是很大的，e1更靠近Ithaca。

他们的标准差为 $\sqrt{s_{1,1}} = 13.62^\circ\mathrm{F}$ $\sqrt{s_{2,2}} = 8.81^\circ\mathrm{F}$

两变量的协方差矩阵 $[S_X] = \begin{bmatrix} 185.47 & 110.84 \\ 110.84 & 77.68 \end{bmatrix}$

两变量协方差阵的两个特征矢量为 $e_1^{\mathrm{T}} = [0.848, 0.530]$ $e_2^{\mathrm{T}} = [-0.530, 0.848]$

特征矢量矩阵$[E] = \begin{bmatrix} 0.848 & -0.530 \\ 0.530 & 0.848 \end{bmatrix}$

特征值为：

$\lambda_1 = 254.76^\circ \mathrm{F}^2$
$\lambda_2 = 8.29^\circ \mathrm{F}^2$
$\lambda_1 + \lambda_2 = s_{1,1} + s_{2,2} = 263.05^\circ \mathrm{F}^2$

- PC1的解释方差非常大；
- PC1体现了区域（两地）最小温度的主要变化特征；
- PC2可以认为是由上述两地构成的区域最小温度偏离整体区域变化特征的局地

变化；

REOF与EOF的对比：

EOF关注全局，不优待某个变量。
EOF正交性限制。
适合压缩数据

REOF简化结构，强调局部
寻找场的物理结构时，RPCA会更理想。
正交性不存在了，方差主导特征不存在了，均匀分给其他变量。

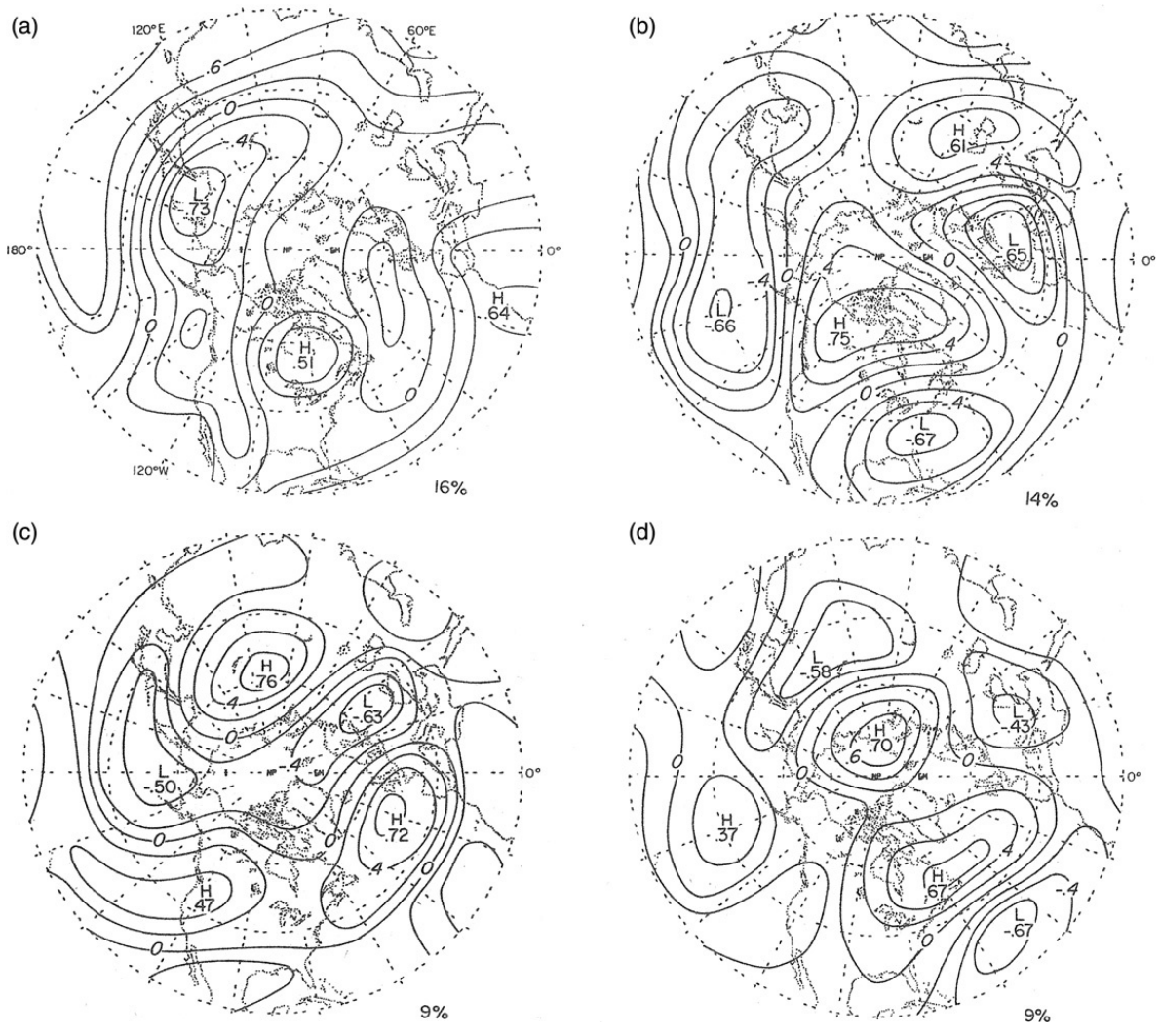


Figure 10.4 This PCA was computed using the correlation matrix of the height data, and scaled so that $\|e_w\| = \lambda_m^{1/2}$, The patterns resemble the teleconnectivity patterns for the same data.

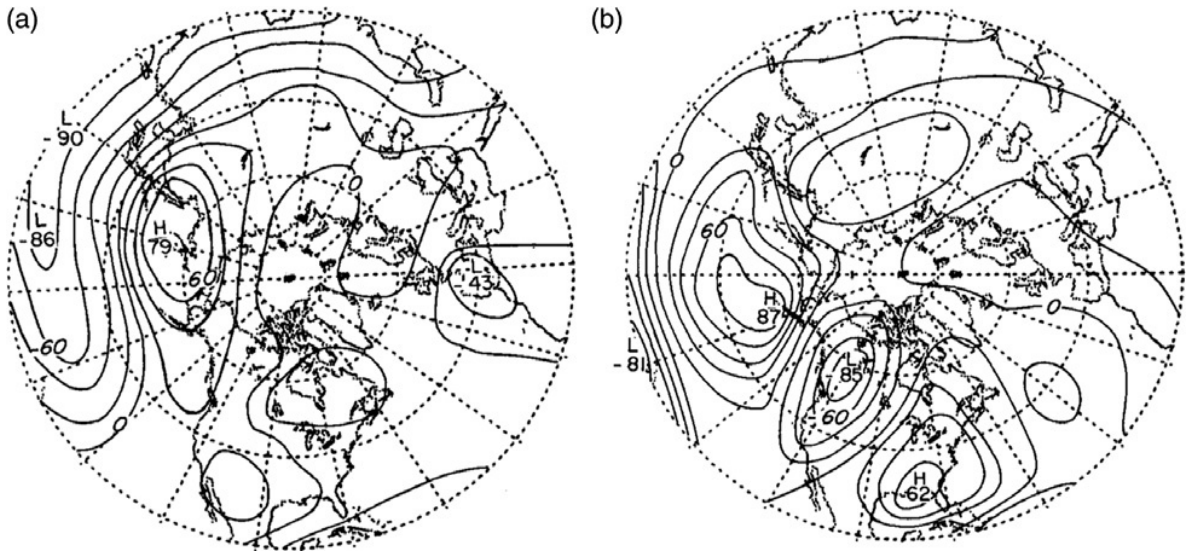


Figure 10.24 Spatial displays of the first two rotated eigenvectors of monthly-averaged hemispheric winter 500-mb heights. The data are the same as those underlying Figure 12.4, but the rotation has better isolated the patterns of variability, allowing a clearer interpretation in terms of the teleconnection patterns in Figure 3.29. From Horel (1981).

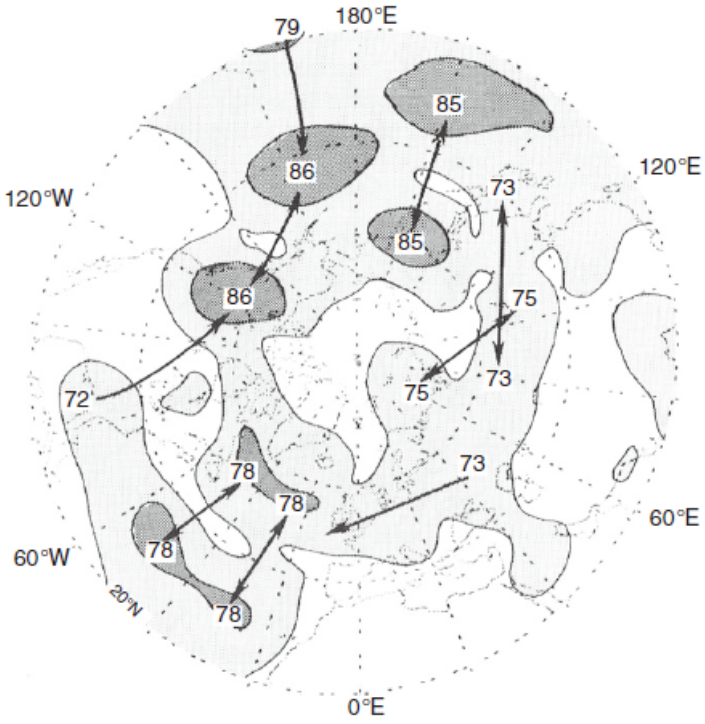


Figure 3.29 Teleconnectivity, or absolute value of the strongest negative correlation from each of many one-point correlation maps plotted at the base gridpoint, for winter 500-mb heights. From Wallace and Blackmon (1983).

It shows the teleconnectivity map for northern hemisphere winter 500-mb heights. The density of the shading indicates the magnitude of the individual gridpoint teleconnectivity values. The locations of local maxima of teleconnectivity are indicated by the positions of the numbers, expressed as x 100. The arrows in Figure 3.29 point from the teleconnection centers (i.e., the local maxima in Ti) to the location with which each maximum negative correlation is exhibited.

第九章 EEMD方法

集合经验模分解：局部自适应时间序列分析技术，适合于分析非线性、非平稳的时间序列。它把复杂数据分解为有限个不同时间尺度的震荡分量。没有实现引入基函数。

第十章 气候极值

气候是所有天气现象的综合表述。C = pdf(Wi)

对于特定的气象要素（如温度或降水），气候就是所有可能天气值所构成的某种概率分布。

气候极值指那些远离气候平均态的小概率的极端事件。

如何构造一个正午气候分布？进而评估其平均和极端状态？

不考虑季节循环的情况下，用24小时逐时资料，365天的的均值，构造一个24小时的温度分布。并算出这些序列的3%和97%分位数，作为极端值。

考虑季节的情况下，构造一个365天12时的序列，用20年的数据来算出每一天正午值的平均，然后算出这些序列的3%和97%分位数，作为极端值。

极值分布GEV

GLM回归模拟
优点：把所有资料纳入一个关于分布（包括均值、极值）的非平稳统计框架，结果具有优越的统计稳定性。
缺点：必须预设某种分布，超拟合现象。

第十一章 气候的非均一性

非均一性：inhomogeneity 气候序列中某些时段由于非自然原因造成的**系统偏差。**（台站迁移、观测规则/仪器改变、卫星更替（如TOPEX/POSEIDON -JASON）资料处理不当）

均一化：检测校订资料的非均一性。

第十二章 随机天气发生器

SWGs（Stochastic Weather Generator） can produce elaborate random numbers which are statistically resemble weather observations, via Monte-Carlo simulations. SWGs are not designed for weather forecasting, but usually for climate studies.