地球惑星科学(地球物理学）特別講義 1 気象データ解析概論
<Special Lecture on Geophysics 1: Outline of Meteorological Data Analysis>

2015 年 12 月 9 日～ 11 日

藤部文昭 (FUJIBE Fumiaki, ffujibe.bs@gmail.com)

Ⅰ 統計の基本 (Basic concepts of statistical analyses)
- 1-1 統計の基本，前置き (Introduction)
- 1-2 確率分布 (Probability distribution)
- 1-3 基本的な統計量 (Basic statistics)
- 1-4 複数変数の統計 (Multivariate statistics)

Ⅱ 最小 2 乗法と最尤法 (Least squares method and the maximum likelihood method)
- 2-1 最小 2 乗法 (Least squares method; LSM)
- 2-2 最尤法 (Maximum likelihood method; MLE)

Ⅲ 二項分布, Poisson 分布，正規分布および中心極限定理 (Binomial distribution, Poisson distribution, normal distribution, and the central limit theorem)
- 3-1. 二項分布 (Binomial distribution)
- 3-2. Poisson 分布 (Poisson distribution)
- 3-3. 正規分布 (Normal distribution, Gauss distribution)
- 3-4. チェビシェフの不等式 (Tchebyshev's inequality)
- 3-5. 中心極限定理 (Central limit theorem)

Ⅳ 正規分布に関連する確率分布 (Distributions derived from the normal distribution)
- 4-1 χ^2 分布 (Chi−square distribution)
- 4-2 *t* 分布 (Student's *t* distribution)
- 4-3 *F* 分布 (*F* distribution)

Ⅴ 主成分分析 (Principal component analysis, PCA)
- 5-1 主成分分析の考え方 (Basic idea)
- 5-2 PCA の定式化 (Formulation)
- 5-3 回転 PCA (Rotated PCA)
- 5-4 主成分分析，経験的直交関数展開, 因子分析の違い (Differences of PCA, empirical orthogonal functions, and factor analysis)

Ⅵ 極値統計 (Extreme value analysis)
- 6-1 極値統計の基本概念 (Basic idea)
- 6-2 極値統計の数学的基礎 (Mathematical basis)
- 6-3 極値統計の手法 (Methods)
- 6-4 極値統計の精度 (Confidence of extreme value analysis)
- 6-5 地域頻度解析 (Regional frequency analysis)
- 6-6 異常値の問題 (Outliers)
- 6-7 閾値解析，POT 解析 (Peaks-over-threshold analysis)

Ⅶ 気象観測とデータ (Meteorological observation and data)

Ⅰ 統計の基本 (Basic concepts of statistical analyses)

1-1 統計の基本，前置き (Introduction)

■ 統計とは

・少数のデータ (標本，sample) から多数の特性 (母集団＝ population) の値を推定する手続き

・例：

世論調査 (public opinion polls)：対象者の回答から，国民 (有権者) 全体の意見を推定する．

気候研究 (climate studies)：ある都市と郊外の有限期間の気温観測結果から，その都市のヒートアイランドの強さを推定する．

■ 母集団と標本

・母集団とは何か

統計学の基本的概念…多数 (無尽蔵) の母集団 (population) から抽出される少数 (有限個) の標本 (sample)

例：サイコロを振る…可能な無限回の試行 (trial)＝母集団，現実の有限回の試行＝標本

・しかし現実には，「無尽蔵の母集団」は必ずしも実在しない．

例：

現在気候における京都の年間の猛暑日数．

気候は長期的に変動するので，「現在気候」の条件が成り立つのはせいぜい数十年程度？

ある稀少な病気における，ある合併症の発症率

その病気の患者が国内に (世界に) 数十人しかいないとしたら･･･それ以上の母集団は実在しない．

・このような場合は，想像上の母集団 (imaginary population) を設定せざるを得ない．

現在の気候状態が何万年も変わらずに続くとしたら･･･

その病気の患者が何万人もいたとしたら･･･

■ データ (標本) の変動と誤差

・データは変動 (variation, variability) →統計結果の不確実性 (uncertainty)

統計とはデータの変動や不確実性を数量的に扱う手続き．

→ 信頼幅 (confidence range), 統計的検定 (statistical test), 有意性 (significance)

変動のないものは統計の対象にならない．「1 週間の日数は平均 7 日だ」？

・シグナルと誤差の概念

シグナル (signal)：データから知りたいもの．

誤差 (error)，ノイズ (noise)：シグナル以外のもの．必ずしも観測ミスや不具合によるものだけではない．

・例：

地球温暖化による長期変化を知りたいとき (detection of global warming)

signal ＝地球温暖化による変化 (trend due to global warming)

noise ＝自然の気候変動 (natural variability)，都市化による変化 (urban warming)，観

測方法の変遷による観測値の変化 (observational bias), ･･･

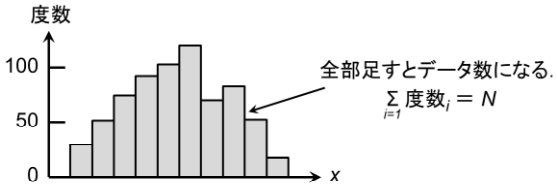
・誤差には 2 種類がある.
ランダム誤差 (random error) : 偶然に支配され (governed by chance), 平均が 0 である誤差 (zero mean).
系統誤差 (systematic error), バイアス (bias) : データに共通に存在し, 平均が 0 でない誤差 (non-zero mean). 何か必然的な要因 (causal factor) が関わる.
統計的手法はランダム誤差の扱いには強いが, バイアスの扱いは不調法.
数学としての統計学における誤差はランダム・単純,
気候統計が扱う誤差 (ノイズ) は複雑・多様
→気候学的知見も使って適切に対処する必要がある.

■ データの独立性 (independence)
・データ同士が無関係, ランダムであること.
・統計解析 (特に, 結果の信頼性評価) に当たって, データの独立性は最重要点の 1 つである.
・しかし気象 (気候) データは, 完全には独立でないことが多いので, 独立性の確保 (独立性が期待できるよう, データを用意 or 処理する) が大きな課題になる.
例 : 日別データ (daily data) 同士は, 必ずしも独立とは言えない (高・低温は数日以上続く, 等).
しかし, ある年の 10 月 1 日, 翌年の 10 月 1 日, 翌々年の 10 月 1 日なら････
・自由度 (degree of freedom): データのうち, 実質的に独立と考えられる組み合わせ数.

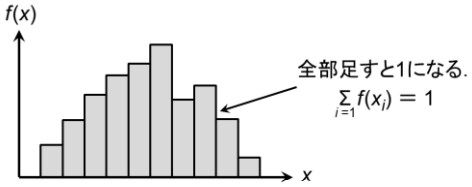
■ データの均質性 (homogeneity)
・データが母集団をうちの一部に偏っていないこと.
・世論調査の場合
特定の世代／生活水準／生活様式･･･に偏っていないか.
電話調査は, 昼間在宅者や固定電話の所有者に偏る?
RDD (Random digit dialing)
・気象 (気候) データの場合
都市への偏り (都市バイアス urban bias)
平地への偏り (山岳地点が少ない)
その他, データごとに観測方法や収録方法を確認すべき.

■ 良い気候統計研究の条件
・データの独立性, 均質性の見極めが的確になされている.
・ランダム誤差による不確実性が適切に評価されている.
・バイアスによる不確実性について, 適切な評価あるいは考察がされている.
・複数の方法を試して結果を比べる. ← 解析の robustness を確認する意味で大事.

1-2 確率分布 (Probability distribution)
■ 度数分布 (frequency distribution), 確率分布 (probability distribution)
・度数分布 : 標本を意識した概念



・確率分布 (probability distribution)：母集団を意識した概念



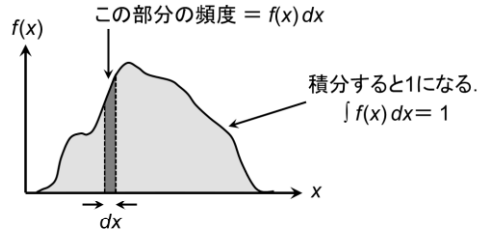
上記は離散分布 (discrete distribution).
変数 (確率変数) が連続値を取るときは，確率密度関数 (probability density function = PDF) を使う.

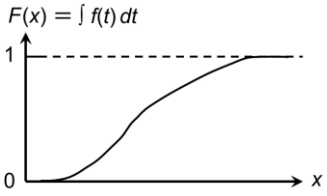
$$\int_{-\infty}^{\infty} f(x)\, dx = 1 \tag{1-1}$$

・累積分布関数 (cumulative distribution function = CDF): 確率分布を積算したもの

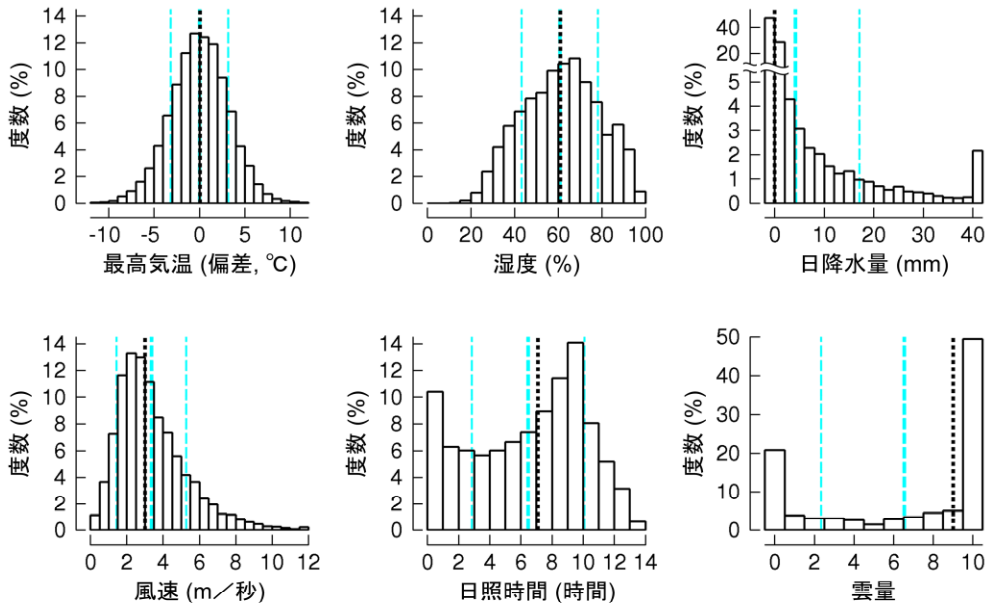
$$F(x) = \int_{-\infty}^{x} f(t)\, dt \tag{1-2}$$

$$f(x) = \frac{dF(x)}{dx} \tag{1-3}$$

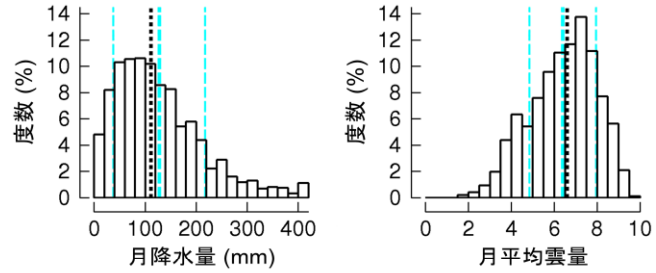




・例: 日別気象要素の度数分布 (東京; 風速・雲量は 09 時の値)



・月別気象要素の度数分布 (東京)



1-3 基本的な統計量 (Basic statistics)

・平均 (mean)

$$\mu = \sum_{i=1} x_i \quad \times \text{度数} / N$$

$$= \sum_{i=1} x_i f_i \tag{1-4}$$

連続分布なら Σ が ∫ になる．すなわち

$$\mu = \int_{-\infty}^{\infty} x f(x)\, dx \tag{1-5}$$

・分散 (variance)

$$\mathrm{var}(x) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx$$

$$= \int_{-\infty}^{\infty} x^2 f(x)\, dx - 2\mu \int_{-\infty}^{\infty} x f(x)\, dx + \mu^2 \int_{-\infty}^{\infty} f(x)\, dx$$

$$= x^2 \text{ の平均 } - \mu^2 \tag{1-6}$$

・標準偏差 (standard variation) ＝ $\sqrt{\mathrm{var}(x)}$ ＝ σ
変動係数 (coefficient of variation = CV) = σ/μ
最低値が 0 である量にのみ意味を持つ (降水量，風速など．気温はダメ).

・モーメント (moment, 積率)

$$\mu'_k = \int_{-\infty}^{\infty} x^k f(x)\, dx \tag{1-7}$$

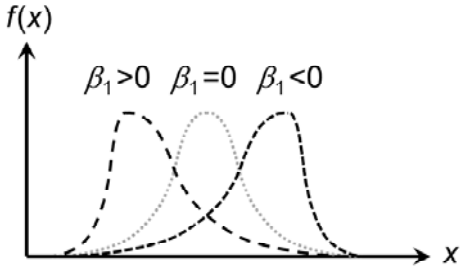
$$\mu_k = \int_{-\infty}^{\infty} (x-\mu)^k f(x)\, dx \tag{1-8}$$

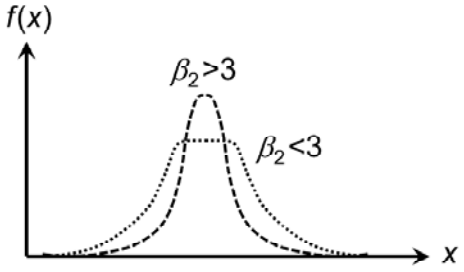
1 次の moment：平均
$\mu = \mu'_1$

2 次の moment：分散
$\mathrm{var}(x) = \mu_2$

3 次以上の moments：分布の偏りを表す.
$\beta_1 = \mu_3^2/\mu_2^3$ ・・・歪度 (skewness)
左右対称な分布なら $\beta_1 = 0$

$\beta_2 = \mu_4/\mu_2^2$ ・・・尖度 (kurtosis, degree of peaking)
正規分布なら $\beta_2 = 3$





・簡単な例
サイコロの目
$f_i = 1/6$ for $i = 1 \sim 6$

$\mu = 3.5$
$\mathrm{var}(i) = 35/12 = 2.917$, $\sigma = 1.708$
$\beta_1 = 0$, $\beta_2 = 303/175 = 1.73$ (< 3)

一様分布

$f(x) = 1$ for $0 \leqq x \leqq 1$

$\mu = 1/2$

$\mathrm{var}(x) = 1/12$, $\sigma = 0.289$

$\beta_1 = 0$, $\beta_2 = 9/5 = 1.8$ (< 3)

■ 順位尺度

・中央値 (median), 分位値 (quantile), パーセンタイル (percentile)

平均や分散に比べ, 使われることは少ない.

検定 (不確実性の評価) が難しい.

← Monte Carlo 的な方法を使えば可能? (bootstrap など)

■ L-moments

moments に対応する 1 次元の尺度 (2 乗以上の項を含まない. L は linear の L)

$$\lambda_1 = E(X_{1:1}) \tag{1-9}$$

$$\lambda_2 = \frac{1}{2} E(X_{2:2} - X_{1:2}) \tag{1-10}$$

$$\lambda_3 = \frac{1}{3} E(X_{3:3} - 2X_{2:3} + X_{1:3}) \tag{1-11}$$

$$\lambda_4 = \frac{1}{4} E(X_{4:4} - 3X_{3:4} + 3X_{2:4} - X_{1:4}) \tag{1-12}$$

ここで, $X_{r:n}$ は, 母集団の中から n 個の標本を取り出したときの r 番目に小さい標本の値を表す.

$\lambda_1 = \mu$

λ_2 は $\mathrm{var}(x)$ に対応. L-CV = λ_2/λ_1

L-moments についての詳細は Hosking and Wallis (1997), 外山・水野 (2002), 三浦・水野 (2005) 参照.

L-moments を計算するプログラム (Fortran 77): http://lib.stat.cmu.edu/general/lmoments

■ より高度な統計量

・期待値 (expected value)

母集団についての平均値のこと.

$$E[h(x)] = \int_{-\infty}^{\infty} h(x)\, f(x) dx \tag{1-13}$$

この表記を使うと

$\mu = E[x]$, $\mathrm{var}(x) = E[(x-\mu)^2]$,
$\mu'_k = E[x^k]$, $\mu_k = E[(x-\mu)^k]$

加法性 (additivity)

$E[h_1(x) + h_2(x)] = E[h_1(x)] + E[h_2(x)]$

・moment generating function (MGF)

$$M_x(t) = E[e^{xt}] = \int_{-\infty}^{\infty} e^{xt} f(x)\, dx \tag{1-14}$$

$$= E\left[1 + tx + \frac{(tx)^2}{2} + \cdots\right] = \sum_{n=0}^{\infty} \frac{1}{n!} \mu'_n t^n \tag{1-15}$$

$$\mu'_n = \left| \frac{\partial^n M_x(t)}{\partial t^n} \right|_{t=0} \tag{1-16}$$

μ のまわりの MGF

$$M_\mu(t) = E[e^{(x-\mu)t}] = e^{-\mu t} M_x(t) \tag{1-17}$$

・特性関数 (characteristic function)

$$\phi_x(t) = E[e^{itx}] = M_x(it) \tag{1-18}$$

特性関数の性質

x と y が independent なら

$$\phi_{x+y}(t) = E[e^{it(x+y)}] = E[e^{itx}]E[e^{ity}] = \phi_x(t)\phi_y(t)$$

一般に $x_1, x_2, \cdots x_n$ が independent なら

$$\phi_{x_1+x_2+\cdots+x_n}(t) = \phi_{x_1}(t)\phi_{x_2}(t)\cdots \phi_{x_n}(t) = \prod_{i=1}^{n} \phi_{x_i}(t) \tag{1-19}$$

変数の 1 次変換 (linear transformation)

$$\phi_{a+bx}(t) = E[e^{it(a+bx)}] = e^{ita} E[e^{itbx}] = e^{ita}\, \phi_x(bt) \tag{1-20}$$

inversion theorem

$\phi_x(t)$ は $f(x)$ のフーリエ変換になっている.

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_x(t)\, e^{-itx} dt \tag{1-21}$$

・random variable の function の density function

random variable: x

density function $f(x)$ が既知のとき, 関数 $y(x)$ の density function は

$y(x)$ が単調関数 $(dy/dx \neq 0)$ なら

$$f[y(x)] = \sum_{\text{all } x} f[x(y)] \left|\frac{dy}{dx}\right|^{-1} \tag{1-22}$$

1-4 複数変数の統計 (Multivariate statistics)

・2 変数の場合

結合確率密度関数 (joint probability density function)

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dxdy = 1 \tag{1-23}$$

x と y が独立 (independent) なら

$$f(x, y) = f(x)\ f^*(y) \tag{1-24}$$

x, y の平均 (mean)

$$\mu_x = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x\ f(x, y)\, dxdy \tag{1-25}$$

共分散 (covariance)

$$\mathrm{cov}(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-\mu_x)(y-\mu_y)\ f(x, y)\, dxdy \tag{1-26}$$

x と y が independent なら $\mathrm{cov}(x, y) = 0$

・多変数の場合

$x_1, x_2, \cdots x_n$ の共分散行列 (covariance matrix)

$$\mathrm{C} = \begin{pmatrix} \mathrm{var}(x_1) & \mathrm{cov}(x_1, x_2) & \cdots & \cdots \\ \mathrm{cov}(x_1, x_2) & \mathrm{var}(x_2) & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \mathrm{var}(x_n) \end{pmatrix} \tag{1-27}$$

■ 相関係数 (correlation coefficient)

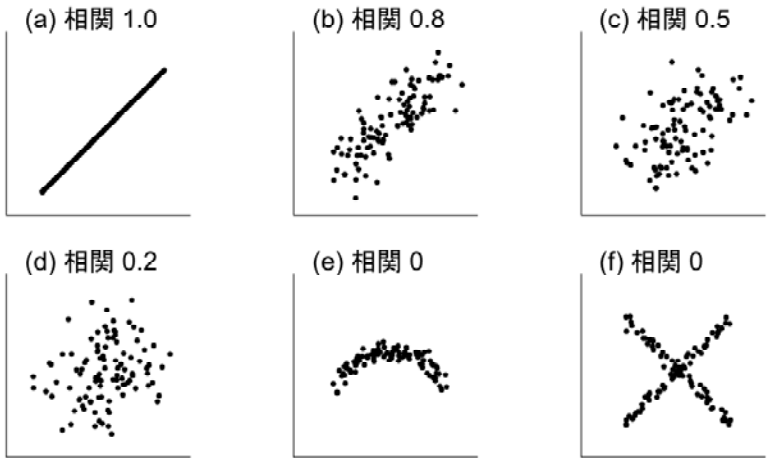
$$\rho(x, y) = \frac{\mathrm{cov}(x, y)}{\sqrt{\mathrm{var}(x)\ \mathrm{var}(y)}} \tag{1-28}$$

correlation coefficient は x と y の線形性の度合い (linearity) を表す.

x と y が線形 (linear) すなわち $y = ax+b$ $(a \neq 0)$ なら $\mathrm{cov}(x, y) = \pm 1$

x と y が independent なら $\rho(x, y) = 0$

しかし, $\rho(x, y) = 0$ でも independent だとは限らない.



・correlation coefficient についての要注意点
相関は，必ずしも因果関係 (causal relationship) を意味しない.
ひと夏のビール販売量と熱中症患者数
小学生の身長と算数の点数

・多変数の場合，$x_1, x_2, \cdots x_n$ の相関行列 (correlation matrix)

$$R = \begin{pmatrix} 1 & \rho(x_1, x_2) & \cdots & \cdots \\ \rho(x_1, x_2) & 1 & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & 1 \end{pmatrix} \tag{1-29}$$

■ 偏相関 (partial correlation)
他の変数の影響を除いた (＝他の変数が同じ値だとしたときの) 相関係数

相関行列の逆行列 $R^{-1} = \begin{pmatrix} r_{11} & r_{12} & \cdots & \cdots \\ r_{12} & r_{22} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & r_{nn} \end{pmatrix}$

において，x_1 と x_2 の偏相関 $\rho_p(x_1, x_2) = -\dfrac{r_{12}}{\sqrt{r_{11}\, r_{22}}}$ (1-30)

http://www.ae.keio.ac.jp/lab/soc/takeuchi/lectures/5_Parcor.pdf

$\rho(x_1, x_2) > 0$ だが $\rho_p(x_1, x_2) < 0$ という場合も珍しくない.

1-5 統計的推定 (Estimation)
測定結果 (＝ある確率分布を持つ population の標本) から
a. PDF および parameter 既知のとき，観測で x を得る確率は?

b. 測定結果 x が与えられたとき，PDF の parameter を求める．

■ 用語の定義

・population の parameter を $\bar{\theta}$ と表記する．

推定値 (estimate): sample $x_1, x_2, \cdots x_n$ による $\bar{\theta}$ の推定値

推定量 (estimator): estimate の population.

一致推定量 (consistent estimator): $P[|\theta_n - \bar{\theta}| < \varepsilon] > 1-\eta$ について，任意の正数 ε, η に対し $n \geqq N$ が存在する．

不偏推定量 (unbiased estimator): $E[\theta_n] = \bar{\theta}$ である推定量.

■ 不偏推定量の例

・標本平均 (sample mean)

$$\bar{x}_n = \bar{x} = \frac{1}{n}\Sigma x_i \tag{1-31}$$

不偏性の証明

$$\mu_{\bar{x}} = E[\frac{1}{n}\Sigma x_i] = \frac{1}{n}\Sigma E[x_i] = \frac{1}{n} n\mu = \mu \tag{1-32}$$

補足：$\bar{x}$の分散は: $\sigma_{\bar{x}}^2 = \sigma/n$

証明：

$$\sigma_{\bar{x}}^2 = E[\{\bar{x}-E(\bar{x})\}^2] = E[(\frac{1}{n}\Sigma x_i - \mu)^2] = \frac{1}{n^2} E[(\Sigma x_i - n\mu)^2] \tag{1-33}$$

右辺を展開すると

$(x_i-\mu)^2$ が n 個: $E[(x_i-\mu)^2] = \int_{-\infty}^{\infty}(x_i-\mu)^2 f(x_i)\, dx_i = \sigma^2$ (1-34)

その他の項は $E[(x_i-\mu)(x_j-\mu)] = \int_{-\infty}^{\infty}(x_i-\mu)(x_j-\mu) f(x_i) f(x_j)\, dx_i\, dx_j$

$$= \int_{-\infty}^{\infty}(x_i-\mu) f(x_i)\, dx_i \int_{-\infty}^{\infty}(x_j-\mu) f(x_j)\, dx_j = 0 \tag{1-35}$$

よって $\sigma_{\bar{x}}^2 = \frac{1}{n^2} n\sigma^2 = \frac{1}{n}\sigma^2$ (1-36)

データが独立なら，分散はデータ数に反比例

標準偏差は$\sqrt{\text{データ数}}$に反比例

・標本分散 (sample variance):

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{1-37}$$

不偏性の証明

$$E[s^2] = \frac{1}{n-1} E[\sum_{i=1}^{n} (x_i - \frac{1}{n} \sum_j x_j)^2]$$

$$= \frac{1}{n-1} E[\frac{n-1}{n} \sum_{i=1}^{n} x_i^2 - \frac{1}{n} \sum_{i \neq j} x_i x_j] = \mu_2' - \mu_1^2 = \sigma^2 \qquad (1\text{-}38)$$

・$\bar{x}$の variance

σ が既知なら，$\sigma_{\bar{x}}^2 = \sigma^2/n$

σが未知なら sample variance $s^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$ を使う．

$\sigma_{\bar{x}}^2 = \frac{1}{n(n-1)} \Sigma (x_i - \bar{x})^2$ しばしば標準誤差 (standard error) と言われる．

■ 信頼区間 (confidence interval)

ある高い割合 (confidence level) でその区間に真値が含まれることを保証する領域

$P[u_1 \leqq u \leqq u_2] = p$

・例: 正規分布

$P[\mu - \sigma \leqq x \leqq \mu + \sigma] = P[\sigma] = 68.3\%$, $P[2\sigma] = 95.4\%$, $P[3\sigma] = 99.7\%$

平均 μ 未知，分散 1 の母集団から sample 100 個を取った場合の $\hat{\mu}$の 95%信頼区間を求める．

$u = \frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}} = 10(\hat{\mu} - \mu)$ は正規分布 $f(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{u^2}{2})$

$P[\bar{\mu} - 0.196 \leqq u \leqq \bar{\mu} + 0.196] = 0.95$

一般に

$$P[\bar{x} - \frac{1.96}{\sqrt{n}} \leqq u \leqq \bar{x} + \frac{1.96}{\sqrt{n}}] = 0.95$$

任意の a に対する $P[\bar{x} - a\sigma \leqq u \leqq \bar{x} + a\sigma]$ の値を求めるには正規分布表を使う．Excel にも実装.

分散が未知なら $\hat{\sigma}^2 = s^2/n = \frac{1}{n(n-1)} \Sigma (x_i - \bar{x})^2$ を使う．

データ数が少ない場合は，正規分布に代えて t 分布を使う．自由度 $\nu = n - 1$.

t 分布表または近似公式を使う．

■ 誤差の伝播 (propagation of error)

$y = y(\theta_1, \theta_2, \cdots \theta_p)$ で，parameter θ_i の error $\Delta\theta_i$ が既知であるとき，y の error Δy はどうなるか.

θ_i の真値 θ_i^* (実際には $\hat{\theta}_i$) について，$(\theta_i - \theta_i^*)$が小なら，

$$y(\theta) = y(\theta^*) + \sum_{i=1}^{p} |(\theta_i - \theta_i^*)\frac{\partial y}{\partial \theta_i}|_{\theta=\theta^*} \tag{1-39}$$

$$\mathrm{var}[y(\theta)] = E[\{y(\theta) - E[y(\theta)]\}^2] \sim E[\{y(\theta) - y(\theta^*)\}^2]$$

$$\sim \sum_{i=1}^{p}\sum_{j=1}^{p} \frac{\partial y}{\partial \theta_i}|_{\theta=\theta^*} \frac{\partial y}{\partial \theta_j}|_{\theta=\theta^*} E[(\theta_i - \theta_i^*)(\theta_j - \theta_j^*)]$$

$$\equiv (\Delta y)^2 \equiv \sum_{i=1}^{p}\sum_{j=1}^{p} \frac{\partial y}{\partial \theta_i}|_{\theta=\theta^*} V_{ij} \frac{\partial y}{\partial \theta_j}|_{\theta=\theta^*} \tag{1-40}$$

無相関なら $V_{ij} = (\Delta\theta_i)^2$ (if $i=j$); 0 (if $i \neq j$) なので

$$(\Delta y)^2 \equiv \sum_{i=1}^{p} (\frac{\partial y}{\partial \theta_i})^2_{\theta=\theta^*} (\Delta\theta_i)^2 \tag{1-41}$$

$V = \sigma^2 W^{-1}$ と書き，W を weight matrix と言う.

Ⅱ 最小 2 乗法と最尤法 (Least squares method and the maximum likelihood method)

2-1 最小 2 乗法 (Least squares method; LSM)

■ LSM の定式化 (formulation)

データに関数を当てはめる方法の 1 つ.

残差平方和を最小にするよう，関数の係数を決める.

・1 次式の場合 (1 次回帰＝ linear regression)

$f(x) = a\,x + b$

$y_i - (a\,x_i + b) = e_i$ として，残差平方和 (residual sum of squares) $S = \Sigma\, e_i^2$ を最小にするように最小 2 乗係数 (least-squares coefficients) a, b を決める.

$$S = \sum_{i=1} [y_i - (a\,x_i + b)]^2 \quad \rightarrow \min. \tag{2-1}$$

解法：

$$\frac{\partial S}{\partial a} = 0, \quad \frac{\partial S}{\partial b} = 0. \tag{2-2}$$

これを展開して連立方程式を解く.

・一般形 (重回帰＝ multiple regression)

$$\eta(x) = \sum_{k=1} \theta_k\, f_k(x) \tag{2-3}$$

$y_i - \eta(x_i) = e_i$ として, 残差平方和 $S = \Sigma\, e_i^2$ を最小にするよう least-squares coefficients θ_k を決める.

$$S = \sum_{i=1} [y_i - \sum_{k=1} \theta_k\, f_k(x)]^2 \quad \rightarrow \min. \tag{2-4}$$

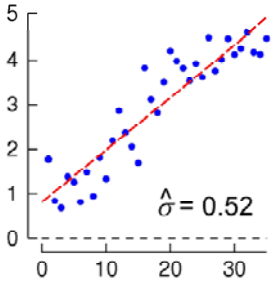
$f_k(x)$は多項式でなくてもいい.　　例：$\cos x$, $\sin x$, $\log x$, ・・・・

階段関数 (step function) $D(x,\ x_0)$：不連続 (discontinuity) の大きさの評価に使える.

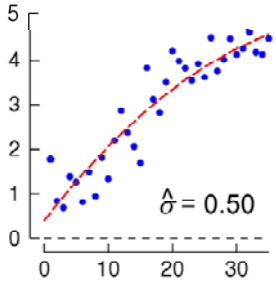
$$D(x, x_0) = 1 \quad \text{if } x \geqq x_0 \tag{2-5}$$
$$D(x, x_0) = 0 \quad \text{if } x < x_0$$

x_0: time of discontinuity (観測所の移転，観測方法の変更など)

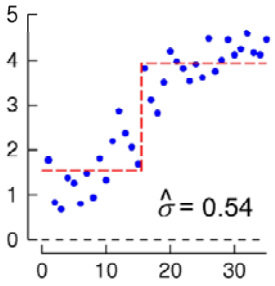
注: x_0 は given でなければならない．x_0 が unknown のときは，この方法は使えない.



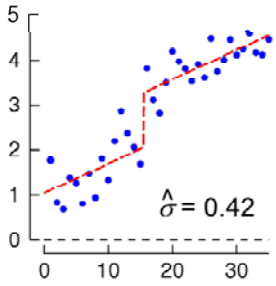
1 次式



2 次式



階段関数 (x_0=15)



1 次式＋階段関数

■ LSM の解法 (solution) :

$$\frac{\partial S}{\partial \theta_k} = 0 \qquad \text{for all } k \tag{2-6}$$

これはベクトルと行列を使って解くのが便利
ベクトルと行列で表示 ($'$ は転置を表す)

$$\mathbf{e} = \mathbf{y} - \mathrm{X}\boldsymbol{\theta} \tag{2-7}$$

ただし

$$\mathrm{X} = \begin{pmatrix} f_1(x_1) & f_2(x_1) & \cdots & f_k(x_1) \\ f_1(x_2) & f_2(x_2) & \cdots & f_k(x_2) \\ \cdots & \cdots & \cdots & \cdots \\ f_1(x_n) & f_2(x_n) & \cdots & f_k(x_n) \end{pmatrix} \qquad n \times k \text{ の行列} \tag{2-8}$$

$$\boldsymbol{\theta}' = (\theta_1,\ \theta_2, \cdots\ \theta_k) \tag{2-9}$$

残差平方和

$$S = \mathbf{e}'\mathbf{e} = (\mathbf{y}-\mathrm{X}\boldsymbol{\theta})'(\mathbf{y}-\mathrm{X}\boldsymbol{\theta}) = \mathbf{y}'\mathbf{y} - 2\boldsymbol{\theta}'\mathrm{X}'\mathbf{y} + \boldsymbol{\theta}'\mathrm{X}'\mathrm{X}\boldsymbol{\theta} \tag{2-10}$$

S を $\boldsymbol{\theta}$ で微分し，0 とする．

$$0 = -2\mathrm{X}'\mathbf{y} + 2\mathrm{X}'\mathrm{X}\boldsymbol{\theta} \ \rightarrow\ (\mathrm{X}'\mathrm{X})\boldsymbol{\theta} = \mathrm{X}'\mathbf{y}$$
$$\rightarrow\ \hat{\boldsymbol{\theta}} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y} \tag{2-11}$$

残差の式

$\mathbf{r} \equiv \mathbf{y} - \boldsymbol{\eta}$　fit した値と observed value の差

$$R = (\mathbf{y}-\mathrm{X}\hat{\boldsymbol{\theta}})'(\mathbf{y}-\mathrm{X}\hat{\boldsymbol{\theta}}) = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathrm{X}\hat{\boldsymbol{\theta}} + \hat{\boldsymbol{\theta}}'\mathrm{X}'\mathrm{X}\hat{\boldsymbol{\theta}} \tag{2-12}$$

これに解を入れると　$R = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathrm{X}\hat{\boldsymbol{\theta}}$

■ $\mathrm{var}(\hat{\boldsymbol{\theta}})$ の推定

$\mathbf{y}$ の covariance matrix

$$\mathrm{V} = \begin{pmatrix} \mathrm{var}(y_1) & \mathrm{cov}(y_1, y_2) & \cdots & \cdots \\ \mathrm{cov}(y_2, y_1) & \mathrm{var}(y_2) & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \mathrm{var}(y_n) \end{pmatrix} \equiv \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \cdots \\ \sigma_{21} & \sigma^2 & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \sigma_n^2 \end{pmatrix} \quad (2\text{-}13)$$

ただし

$$\sigma_i^2 = E[e_i^2] = \mathrm{var}(y_i) \quad (2\text{-}14)$$

$$\sigma_{ij} = E[e_i e_j] = \mathrm{cov}(y_i y_j) = \int \cdots \int (y_i - \mu_i)(y_j - \mu_j) f(y_1, y_2, \cdots y_n)\, dy_1 dy_2 \cdots dy_n \quad (2\text{-}15)$$

$\{y_i\}$は互いに独立で，分散は等しいと仮定すれば，$\mathrm{var}(y_i) = \sigma^2$, $\mathrm{cov}(y_i, y_j) = 0$ で

$$\mathrm{V} = \sigma^2 \mathrm{I}_{n\times n} \quad (2\text{-}16)$$

$$\mathrm{var}(\hat{\boldsymbol{\theta}}) = \mathrm{var}\{(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}\}$$
$$= (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{var}(\mathbf{y})\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1} = \sigma^2(\mathrm{X}'\mathrm{X})^{-1} \quad (2\text{-}17)$$

・σ の estimation

$$E[R] = (n-k)\sigma^2 \;\rightarrow\; \hat{\sigma}^2 = \frac{R}{n-k} \quad (2\text{-}18)$$

上式の導出

$$R = (\mathbf{y}-\mathrm{X}\hat{\boldsymbol{\theta}})'(\mathbf{y}-\mathrm{X}\hat{\boldsymbol{\theta}}) = [(\mathbf{y}-\mathrm{X}\boldsymbol{\theta}) - (\mathrm{X}\hat{\boldsymbol{\theta}}-\mathrm{X}\boldsymbol{\theta})]'(\mathbf{y}-\mathrm{X}\hat{\boldsymbol{\theta}})$$
$$= (\mathbf{y}-\mathrm{X}\boldsymbol{\theta})'(\mathbf{y}-\mathrm{X}\hat{\boldsymbol{\theta}}) - (\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})'\underbrace{(\mathrm{X}'\mathbf{y}-\mathrm{X}'\mathrm{X}\hat{\boldsymbol{\theta}})}_{=0}$$

$$= (\mathbf{y}-\mathrm{X}\boldsymbol{\theta})'(\mathbf{y}-\mathrm{X}\boldsymbol{\theta}) - (\mathbf{y}-\mathrm{X}\boldsymbol{\theta})'\mathrm{X}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}) \quad (2\text{-}19)$$

これの第 1 項については

$$E[(\mathbf{y}-\mathrm{X}\boldsymbol{\theta})'(\mathbf{y}-\mathrm{X}\boldsymbol{\theta})] = n\sigma^2 \quad (2\text{-}20)$$

第 2 項については,

$\hat{\boldsymbol{\theta}} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}$ により

$$\hat{\boldsymbol{\theta}}-\boldsymbol{\theta} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y} - \boldsymbol{\theta}$$
$$= (\mathrm{X}'\mathrm{X})^{-1}(\mathrm{X}'\mathbf{y}-\mathrm{X}'\mathrm{X}\boldsymbol{\theta}) = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'(\mathbf{y}-\mathrm{X}\boldsymbol{\theta})$$

となるので

$$E[(\mathbf{y}-\mathrm{X}\boldsymbol{\theta})'\mathrm{X}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})] = E[(\mathbf{y}-\mathrm{X}\boldsymbol{\theta})'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'(\mathbf{y}-\mathrm{X}\boldsymbol{\theta})] \quad (2\text{-}21)$$

一般に $E[\mathbf{x}'\mathrm{C}\mathbf{x}] = E[\sum_{i,j} x_i C_{ij} x_j]$.

$E[\Sigma\, x_i x_j] = \sigma^2$ (if $i=j$), 0 (if $i \neq j$) なら,
$E[\mathbf{x}'\mathrm{C}\mathbf{x}] = E[\sum_i x_i^2 C_{ii}] = \sigma^2 \mathrm{tr}(\mathrm{C})$.

よって(2-21)は

$$E[(\mathbf{y}-\mathrm{X}\boldsymbol{\theta})'\mathrm{X}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})] = \mathrm{tr}\,[\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}']\,\sigma^2$$
$$= \mathrm{tr}\,[(\mathrm{X}'\mathrm{X})(\mathrm{X}'\mathrm{X})^{-1}]\,\sigma^2 = \mathrm{tr}\,[\mathrm{I}_{k\times k}]\,\sigma^2 = k\sigma^2 \qquad (2\text{-}22)$$

以上から

$$\mathrm{var}(\hat{\boldsymbol{\theta}}) = \frac{R}{n-k}(\mathrm{X}'\mathrm{X})^{-1} \equiv \mathrm{U}\ (\text{error matrix}) \qquad (2\text{-}23)$$

$$\hat{\theta}_i \text{ の error } \Delta\hat{\theta}_i = \sqrt{U_{ii}} \qquad (2\text{-}24)$$

■ 補足

(1) 回帰直線への距離の 2 乗和を最小にする方法 (perpendicular offsets)

$$S = \sum_{i=1}^{N} \frac{[y_i - (a x_i + b)]^2}{a^2 + 1} \rightarrow \min., \qquad (2\text{-}25)$$

$\partial S/\partial a = 0$, $\partial S/\partial b = 0$ を解くと,

$$a = c \pm\sqrt{c^2+1}$$
$$b = (Q - aP)/N \qquad (2\text{-}26)$$

ただし $P = \Sigma\, x_i$, $Q = \Sigma\, y_i$, $R = \Sigma\, x_i^2$, $S = \Sigma\, y_i^2$, $T = \Sigma\, x_i y_i$ で

$$c = \frac{P^2 - Q^2 + N(S-R)}{2(NT - PQ)}, \qquad (2\text{-}27)$$

であり, a の式の複号は(2-25)の右辺が最小になる方をとる.

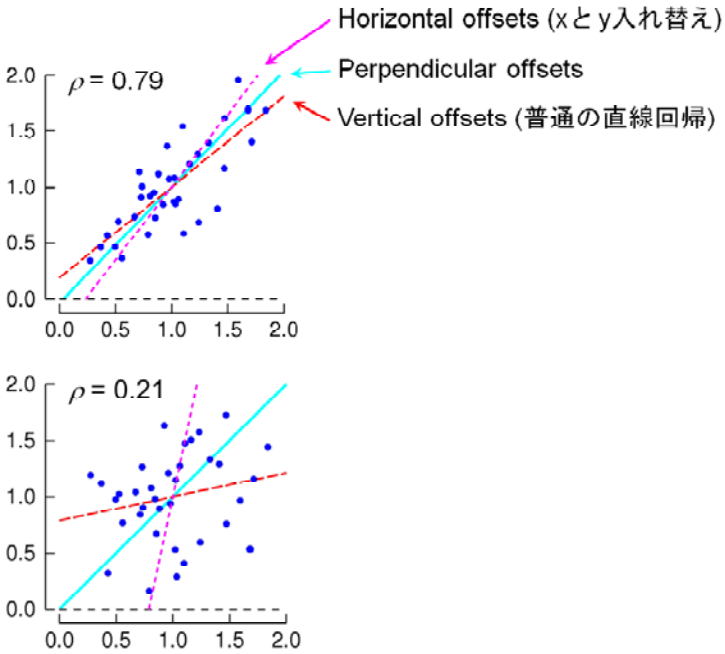
この方法を使うには, 縦軸と横軸の量が同質である必要がある.
例：親の身長と子の身長；大阪の気温と京都の気温

(2) linear regression の場合, x と y を逆にした場合の回帰直線は, 元の回帰直線とは一致しない.
例：x が親の身長, y が子の身長のとき
y に $(ax+b)$ を当てはめると, 一般に $a < 1$.
(親が極端に高身長や低身長でも, 子の身長はこれほど極端ではない傾向がある･･･「回帰」の語源)
一方, x に $(a'y+b')$ を当てはめても, $a' < 1$ になる.
(子が極端に高身長や低身長でも, 親の身長はこれほど極端ではない傾向がある)

同様に,
x と y に相関がある場合, 回帰直線は相関の“主軸”とは一致しない (相関の主軸よ

りも傾きが小さくなる)

この問題は気候研究の落とし穴になり得る．例えば x を大阪の気温，y を京都の気温として y に $(ax+b)$ を当てはめたとき，$a < 1$ であっても「京都の気温は大阪の気温より変動幅が小さい」と結論するのは誤りである．



(3)

最小 2 乗法では，誤差の確率分布についての仮定はない．

ただし，

・係数の信頼幅を求める際には正規分布の仮定が使われる．

・誤差が正規分布なら，最尤法と一致する．

(4) 非線形 LSM

当てはめる関数が，係数 θ_k の 1 次関数でない場合

例：$a\cos(bx+c)$ を当てはめたい $(a, b, c$ を最小 2 乗法で求めたい$)$ とき

数値的に解く必要がある．

■ 重みつき最小 2 乗法 (weighted least-squares method)

$$\mathrm{V} = \begin{pmatrix} \sigma_1^2 & \cdots & \sigma_{1n} \\ \cdots & \cdots & \cdots \\ & & \sigma_n^2 \end{pmatrix} = \sigma^2 \mathrm{W}^{-1} \quad \mathrm{W} \text{ は weight matrix} \tag{2-28}$$

$$S = (\mathbf{y}-\mathrm{X}\boldsymbol{\theta})'\mathrm{W}(\mathbf{y}-\mathrm{X}\boldsymbol{\theta})/\sigma^2 \text{ を最小にする．} \tag{2-29}$$

$$\hat{\boldsymbol{\theta}} = (\mathrm{X}'\mathrm{W}\mathrm{X})^{-1}\mathrm{X}'\mathrm{W}\mathbf{y},\ \mathrm{var}(\hat{\boldsymbol{\theta}}) = \sigma^2(\mathrm{X}'\mathrm{W}\mathrm{X})^{-1} \tag{2-30}$$

$$\hat{\sigma}^2 = (\mathbf{y}-\mathrm{X}\hat{\boldsymbol{\theta}})'\mathrm{W}(\mathbf{y}-\mathrm{X}\hat{\boldsymbol{\theta}})/(n-k) \tag{2-31}$$

$$\text{error matrix } \mathrm{U} = (\mathbf{y}-\mathrm{X}\boldsymbol{\theta})'\mathrm{W}(\mathbf{y}-\mathrm{X}\boldsymbol{\theta})(\mathrm{X}'\mathrm{W}\mathrm{X})^{-1}/(n-k) \tag{2-32}$$

2-2 最尤法 (Maximum likelihood method; MLE)

最も一般的な parameter estimation の方法.

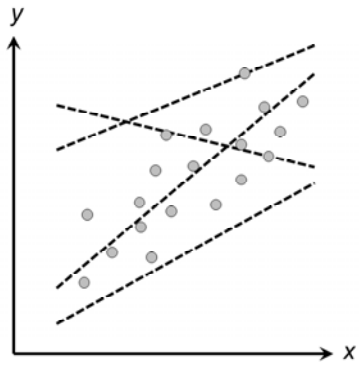
■ 最尤法の考え方

・例 1: 10 人に内閣支持率を尋ねたところ,「支持」が 3 人だった.
このとき真の (＝母集団の) 内閣支持率はいくらか？

答. 30%ぐらいの可能性が高そう. ･･･しかし, なぜそう思えるのだろう？

次のうち, どれの可能性が一番高そうか？
・真の支持率が 10%のとき, 10 人のうち 3 人が「支持」と答える確率
・真の支持率が 30%のとき, 10 人のうち 3 人が「支持」と答える確率
・真の支持率が 50%のとき, 10 人のうち 3 人が「支持」と答える確率

・例 2: データに関数を当てはめる.
真の分布は直線になることが, 分かっているとする.
データが図のようになっているとき, 真の分布はどの直線である可能性が高そうか.



・尤度の概念
「真の支持率が p であるとき, n 人中 x 人が「支持」と答える確率 (probability)」
← これを,「n 人中 x 人が「支持」と答えたとき, 真の支持率が p である尤度 (likelihood)」と読み替える.
その値は, 二項分布により

$$L = \frac{n!}{x!\,(n-x)!}\,p^x(1-p)^{n-x} \tag{2-33}$$

尤度 L を, p の関数と見なして「尤度関数 (likelihood function)」と言う.
一般に, パラメーター θ を持つ確率密度関数 PDF $f(x;\,\theta)$ を,
x ではなく (x が与えられたときの) θ の関数と見なしたものが尤度関数である.

n=10, x=3 の場合, 真の支持率が p である尤度は
p=0.1 なら L=0.06
p=0.2 なら L=0.20
p=0.3 なら L=0.27 ← 最大尤度 (maximum likelihood)

p=0.4 なら L=0.21
p=0.5 なら L=0.12
p=0.6 なら L=0.04

■ 最尤法の解法

(1) 内閣支持率の問題

式(2-33)の L を最大にする p を計算すればいい.

(2) 直線のあてはめの問題

真の分布が直線 $y=ax+b$ であるとき，データ i の値が y_i である確率
＝ データ i の値が y_i であるとき，真の分布が $y=ax+b$ である尤度は,
誤差 (＝データと真の分布の差) が確率分布 g に従うとすると,

$$L(x_i, y_i; a, b) = g[y_i - (ax_i + b)] \tag{2-34}$$

N 個のデータ i=1 ～ N に対しては，真の分布が $y = ax + b$ である尤度は上記の積になる.

$$L(x_1, \text{---}, x_N; y_1, \text{---}, y_N; a, b) = \prod_{i=1}^{N} g[y_i - (ax_i + b)] \tag{2-35}$$

a, b の求め方: L を最大にする a, b を計算すればいい.

$$\frac{\partial}{\partial a}\log L = 0, \quad \frac{\partial}{\partial b}\log L = 0. \tag{2-36}$$

g が平均 0，分散 σ^2 の正規分布なら

$$g = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\{-\frac{1}{2}\frac{[y-(ax+b)]^2}{\sigma^2}\} \tag{2-37}$$

よって,

$$\log L = \text{定数} - \sum_{i=1}[y_i - (ax_i + b)]^2 \,/\, 2\sigma^2 \tag{2-38}$$

となり， $\Sigma[y_i - (ax_i + b)]^2$ を最小にする最小 2 乗法と同じになる.

上記で，データごとに σ が異なる場合は･･･
データ i の分散を σ_i とすると,

$$\log L = \text{定数} - \sum_{i=1}[y_i - (ax_i + b)]^2 \,/\, 2\sigma_i^2 \tag{2-39}$$

となり，重みつき最小 2 乗法と同じになる.

(3) 一般の場合

$f(x;\overline{\theta})$: PDF, $\overline{\theta}$ は未知の parameter, $x_1, x_2, \cdots x_n$: independent random variable とする.

joint density function $L(x_1, x_2, \cdots x_n; \overline{\theta}) = \prod_{i=1} f(x_i; \overline{\theta})$ について

最尤推定量 (max. likelihood estimator) $\hat{\theta}$は $L(\theta)$を最大にするもの．よって

$$\frac{\partial}{\partial\theta}\log L(\theta) = 0 \tag{2-40}$$

■ $\hat{\theta}$の variance の推定

$L(\theta)$は random variable θの density function と見なせるので,

$$\mathrm{var}(\theta) = \frac{\int_{-\infty}^{\infty}(\theta-\hat{\theta})^2 L(\theta)\,d\theta}{\int_{-\infty}^{\infty} L(\theta)\,d\theta} \tag{2-41}$$

一方，n が大なら中心極限定理 (3.5) により,

$$L(\theta) = \frac{1}{\sqrt{2\pi V}}\exp[-\frac{1}{2}\frac{(\theta-\hat{\theta})^2}{V}] \qquad \text{ただし } V = \mathrm{var}(\theta) \tag{2-42}$$

よって

$$\log L(\theta) = -\log\sqrt{2\pi V} - \frac{1}{2}\frac{(\theta-\hat{\theta})^2}{V} \tag{2-43}$$

$$\frac{\partial^2}{\partial\theta^2}\log L(\theta) = -\frac{1}{V} \quad \text{よって} \quad \mathrm{var}(\hat{\theta}) = [-\frac{\partial^2}{\partial\theta^2}\log L(\theta)]^{-1}|_{\theta=\hat{\theta}} \tag{2-44}$$

・例 1: 正規分布

$$f(x;\,\mu,\,\sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}] \tag{2-45}$$

について，データ $x_1, \cdots, x_n$ が得られたとき,
μ および $\mathrm{var}(\mu)$を推定する (σは既知とする).

$$\log L(\mu) = -n\log[\sqrt{2\pi}\,\sigma] - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \tag{2-46}$$

μ の最尤推定量は

$$\frac{\partial\log L(\mu)}{\partial\mu} = 0 \rightarrow \quad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}x_i = \bar{x} \tag{2-47}$$

$$\mathrm{var}(\hat{\mu}) = [-\frac{1}{2\sigma^2}\Sigma\frac{\partial^2}{\partial\mu^2}(x_i-\mu)^2]^{-1}|_{\mu=\hat{\mu}} = \frac{\sigma^2}{n} \tag{2-48}$$

・例 2: 正規分布する母集団の測定値 x_i が誤差 Δx_i を持つとき，$\hat{\mu}$, $\mathrm{var}(\hat{\mu})$を求める (σ 未知).

$$f(x;\ \mu;\ \Delta x) = \frac{1}{\sqrt{2\pi}\,\Delta x^2}\exp[-\frac{1}{2}(\frac{x-\mu}{\Delta x})^2] \tag{2-49}$$

$$\hat{\mu} = \frac{\sum_{i=1}^{n}\frac{x_i}{\Delta x_i^2}}{\sum_{i=1}^{n}\frac{1}{\Delta x_i^2}} \qquad \leftarrow \text{重み付き平均 (weighted mean)} \tag{2-50}$$

$$\mathrm{var}(\hat{\mu}) = [\frac{\partial^2 \log L(\mu)}{\partial \mu^2}]^{-1}|_{\mu=\hat{\mu}} = [\sum_{i=1}^{n}(\frac{1}{\Delta x_i})^2]^{-1} \tag{2-51}$$

■ 最尤法の得失

メリット: 確率分布が何でもいい

デメリット: 確率分布が known (given) でないといけない

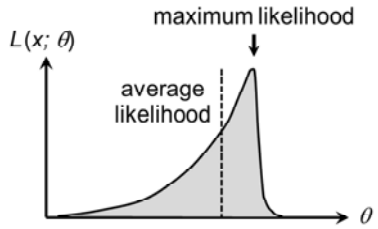
計算が複雑 (多くの場合，数値解法が必要．複雑な確率分布関数だと，log$L$ の微分を計算するだけで一仕事)

・補足:

最尤推定量は unbiased ではない.

尤度関数 $L(\theta)$ が非対称なら,

最大尤度 (maximum likelihood) ≠ 平均尤度 (average likelihood)



Ⅲ 二項分布, Poisson 分布, 正規分布および中心極限定理 (Binomial distribution, Poisson distribution, normal distribution, and the central limit theorem)

3-1. 二項分布 (Binomial distribution)

■ 定義

事象 (event) A の起こる確率 p, 起こらない確率 $q = 1-p$

n 回の試行 (trials) で event A が起きる数 x の PDF

$$f(x) = \frac{n!}{x!\,(n-x)!}\,p^x q^{n-x} \tag{3-1}$$

なお

$$\sum_{x=0}^{n} f(x) = 1 \tag{3-2}$$

理由：

一般に, 任意の p, q について

$$(p+q)^n = \sum_{x=0}^{n} \frac{n!}{x!\,(n-x)!}\,p^x q^{n-x} \tag{3-3}$$

今の場合 $p+q=1$ なので, 上式=1 になる.

■ 平均と分散

$$\mu = \sum_{x=0}^{n} x f(x) = \sum_{x=0}^{n} x\,\frac{n!}{x!\,(n-x)!}\,p^x q^{n-x}$$

$$= np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!\,(n-x)!} p^{x-1} q^{n-x} \qquad y = x-1 \text{ と置く}$$

$$= np \sum_{y=0}^{n-1} \frac{(n-1)!}{y!\,(n-1-y)!} p^y q^{n-1-y} = np \tag{3-4}$$

$$\mu'_2 = \sum_{x=0}^{n} x^2\,\frac{n!}{x!\,(n-x)!} p^x q^{n-x} \qquad x^2 = x(x-1) + x \text{ と置く}$$

$$= \sum_{x=0}^{n} x(x-1)\,\frac{n!}{x!\,(n-x)!} p^x q^{n-x} + np \qquad y = x-2 \text{ と置く}$$

$$= n(n-1)p^2 \sum_{y=0}^{n-2} \frac{(n-2)!}{y!\,(n-2-y)!}\,p^y q^{n-2-y} + np$$

$$= n(n-1)p^2 + np$$

$$\mathrm{var}(x) = \mu'_2 - \mu^2 = np(1-p) = npq \tag{3-5}$$

MGF

$$M_x(t) = \sum_{x=0}^{n} \frac{n!}{x!\,(n-x)!}\, p^x q^{n-x}\, e^{xt} = (pe^t+q)^n \tag{3-6}$$

二項分布 → Poisson 分布 (n が大で $\lambda = np$ が有限の場合)
二項分布 → 正規分布 (n が大)

例：世論調査の結果は，統計上どの程度信頼できるか?
真の内閣支持率が p であるとき，無作為に n 人を選んだとすると
支持数の期待値: $\mu = np$, 支持率の期待値: $\mu/n = p$
支持数の標準偏差: $\sqrt{\mu_2} = \sqrt{npq}$, 支持率の標準偏差: $\sqrt{\mu_2}/n = \sqrt{pq/n}$

真の内閣支持率が p = 50%であるとき，支持率の標準偏差 $= \sqrt{pq/n} = 1/2\sqrt{n}$
対象者数が n =10 人なら，15.8%
対象者数が 100 人なら，5.0%
対象者数が 1000 人なら，支持率の標準偏差 = 1.6%
→ 1 ～ 2%の変動は偶然の範囲内.
対象者数が 100000 人なら，支持率の標準偏差 = 0.16%
→ 有権者数 10 万の選挙区が複数あるとき，選挙区ごとの政治意識に差がなければ，選挙結果 (得票率) の違いは 0.数%以内におさまるはず.

3-2. Poisson 分布 (Poisson distribution)

■ 定義

大量の試行により，発現頻度のごく小さい事象が有限回起きる場合
二項分布からの導出

$$f(x) = \frac{n!}{x!\,(n-x)!} p^x q^{n-x}$$

$$= \frac{n(n-1)\cdots(n-x+1)\,(np)^x(1-p)^{n-x}}{n^x x!}$$

$$= \frac{1(1-1/n)(1-2/n)\cdots[1-(x-1)/n]\,(np)^x(1-p)^n}{(1-p)^x x!} \tag{3-7}$$

$n \to \infty$, $p \to 0$,

$$(1-p)^n = [(1-p)^{-1/p}]^{-np} \to e^{-np} \quad [\because \lim_{z\to 0}(1+z)^{1/z} = e] \tag{3-8}$$

$$\frac{1(1-1/n)(1-2/n)\cdots[1-(x-1)/n]}{(1-p)^x} \to 1 \tag{3-9}$$

従って, $np=\lambda$ (有限＝平均発生回数) とすると

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{3-10}$$

■ 平均と分散

$$\mu = \sum_{x=0}^{\infty} x \frac{\lambda^x e^{-\lambda}}{x!}$$

$$= e^{-\lambda}\lambda \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} = e^{-\lambda}\lambda e^{\lambda} = \lambda \tag{3-11}$$

$$\mu'_2 = \lambda^2 + \lambda; \quad \mathrm{var}(x) = \mu'_2 - \mu^2 = \lambda \tag{3-12}$$

MGF

$$M_x(t) = \sum_{x=0}^{\infty} e^{xt} \frac{e^{-\mu}\mu^x}{x!} = e^{-\mu} \sum_{x=0}^{\infty} \frac{(\mu e^t)^x}{x!} = e^{-\mu} \exp(\mu e^t) \tag{3-13}$$

二項分布との違い: n や p が explicit に出てこない.

適用例: 原子核の崩壊，交通事故の発生件数，極端現象の発生回数など
変化の有意性の有無を大ざっぱに評価する.

豪雨回数 (100mm/h とか) が 40 回から 50 回になった．この変化は本物か?

→ Poisson 分布を考えると，標準偏差は$\sqrt{40} \approx 6.5$ 回，95%信頼幅はその 2 倍 ≈13 回，とすると･･･

3-3. 正規分布 (Normal distribution, Gauss distribution)

■ 定義

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(x-x_0)^2}{2\sigma^2}\right] \tag{3-14}$$

二項分布で n が大きいとき，平均値に近いところの分布は正規分布になる.
導出：

Stirling の式 $x! \approx \sqrt{2\pi}\, x^{x+1/2} e^{-x}$ を使って二項分布の式を書き改め，変形していく.
文献参照 (例えば http://teenaka.at.webry.info/201207/article_19.html)

■ 平均と分散

$$\mu = \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{\infty} x \exp\left[-\frac{(x-x_0)^2}{2\sigma^2}\right] dx \qquad [t = (x-x_0)/\sigma \text{ とする}]$$

$$= \frac{1}{\sqrt{2\pi}} \left\{ \int_{-\infty}^{\infty} \sigma t \exp\left[-\left(\frac{t^2}{2}\right)\right] dt + \int_{-\infty}^{\infty} x_0 \exp\left[-\left(\frac{t^2}{2}\right)\right] dt \right\}$$

$$= 0 + x_0 = x_0 \tag{3-15}$$

$$\mu'_2 = \frac{1}{\sqrt{2\pi}} \left\{ \int_{-\infty}^{\infty} \sigma^2 t^2 \exp\left[-\left(\frac{t^2}{2}\right)\right] dt + \int_{-\infty}^{\infty} 2x_0 \sigma t \exp\left[-\left(\frac{t^2}{2}\right)\right] dt + \int_{-\infty}^{\infty} x_0^2 \exp\left[-\left(\frac{t^2}{2}\right)\right] dt \right\}$$

$$= \sigma^2 + x_0^2 \tag{3-16}$$

$$\mathrm{var}(x) = \mu'_2 - \mu^2 = \sigma^2 \tag{3-17}$$

MGF, CF

$$M_x(t) = E[\exp(tx)] = \exp(\mu t + \sigma^2 t^2/2) \tag{3-18}$$
$$\phi_x(t) = E[\exp(itx)] = \exp(i\mu t - \sigma^2 t^2/2) \tag{3-19}$$

$\mu=0$, $\sigma=1$ なら，標準正規分布 (standard normal distribution)

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}) \tag{3-20}$$

■ 正規分布の性質

(1) 多くの PDF の limitimg form
(2) 多くの物理測定量の分布を表す
(3) random error を含む測定値は真値のまわりに正規分布をなす

3-4. チェビシェフの不等式 (Tchebyshev's inequality)

■ 定義

平均 μ, 分散 σ^2 の任意の分布について，k を任意の正数とすると，

$$P[|x-\mu| \geqq k\sigma] \leqq \frac{1}{k^2} \qquad P[\ \ ] \text{は確率を表す.} \tag{3-21}$$

証明

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx$$

$$\geqq \int_{|x-\mu| \geqq k\sigma} (x-\mu)^2 f(x)\, dx$$

$$\geqq \int_{|x-\mu| \geqq k\sigma} (k\sigma)^2 f(x)\, dx = (k\sigma)^2 \int_{|x-\mu| \geqq k\sigma} f(x)\, dx$$

$$= (k\sigma)^2 P[|x-\mu| \geqq k\sigma] \tag{3-22}$$

チェビシェフの不等式による制約は，現実の確率分布よりもはるかに緩い.

例：正規分布で標準偏差の 2 倍以上の値が出る確率は 4.5%, 3 倍以上の値が出る確率は 0.27%

⟷チェビシェフの不等式では，2 倍以上の値が出る確率は 1/4 = 25%以下，3 倍以上は 1/9 = 11%以下.

しかし，「絶対にあり得ない限界」を与える点で，チェビシェフの不等式は使える.

例：25 人のテストの得点が最高 100 点，最低 0 点であるとき，標準偏差は最低でも 10 点.

∵ σ < 10 点だと，$|x-\mu| > 5\sigma$ の学生が 25 人中最低 1 人は居ることになり，$P[|x-\mu| > 5\sigma] < 1/5^2 = 1/25$ に反する.

■ 大数の弱法則 (weak law of large numbers)

チェビシェフの不等式を大きさ n の標本の平均 $\bar{x}_n$ について書くと，

$$P[|\bar{x}_n-\mu| \geqq \frac{k\sigma}{\sqrt{n}}] \leqq \frac{1}{k^2} \tag{3-23}$$

ε, δ ($\varepsilon>0$, $0<\delta<1$) に対し, $k=\delta^{-1/2}$, $n=\sigma^2/\delta\varepsilon^2$ と置けば,
$m \geqq \sigma^2/\delta\varepsilon^2$ となる整数について

$$P[|\bar{x}_m-\mu| \geqq \varepsilon] \leqq \delta \tag{3-24}$$

$\delta \to 0$, $m \to \infty$の極限で，任意の $\varepsilon>0$ に対し

$$P[|\bar{x}_m-\mu| \geqq \varepsilon] \to 0 \tag{3-25}$$

すなわち，標本数が十分多ければ，標本平均は母平均に一致する．

3-5. 中心極限定理 (Central limit theorem)

■ 定義

平均 μ, 分散 σ^2 を持つ分布関数未知の確率変数を x_i (互いに独立) とすると，標本平均 $\bar{x}_n$ の分布は，n が大きくなると，平均 μ, 分散 σ^2/n の正規分布に近づく．
すなわち，正規分布

$$u(t) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{t^2}{2}) \tag{3-26}$$

について，任意の t_1, t_2 に対し

$$\lim_{n\to\infty} P[t_1 \leqq \frac{\bar{x}_n-\mu}{(\sigma/\sqrt{n})} \leqq t_2] = \int_{t_1}^{t_2} u(t)\,dt \tag{3-27}$$

■ 証明

$s=\sum_{i=1}^{n} x_i$ の平均は $\mu_s=n\mu$, 分散は $\sigma_s^2=n\sigma^2$

$u=(s-\mu_s)/\sigma_s$ を考えると，$u=\frac{1}{\sqrt{n}\sigma}\Sigma(x_i-\mu)$

$x_i-\mu$ の特性関数を ϕ_{x_i} とすると，u の特性関数は

$$\phi_u(t) = \prod_{i=1}^{n} \phi_{x_i}(\frac{t}{\sqrt{n}\sigma}) = [\phi_x(\frac{t}{\sqrt{n}\sigma})]^n = [1+\sum_{r=1}^{\infty} \mu_r' \frac{1}{r!}(\frac{it}{\sqrt{n}\sigma})^r]^n \tag{3-28}$$

$x_i-\mu$ の 1 次 moment= 0, 2 次 moment=σ^2 であることから,

$$\phi_u(t) = [1-\frac{t^2}{2n}+\cdots]^n \to \exp(-\frac{t^2}{2}) \qquad \because \lim_{z\to\infty}(1+\frac{1}{z})^z \to e \tag{3-29}$$

inversion theorem (1-21) により,
s は平均 μ_s, 分散は σ_s^2 の正規分布になる．
よって, $\bar{x}_n$ は平均 μ, 分散 σ^2/n の正規分布になる．

Ⅳ 正規分布に関連する確率分布 (Distributions derived from the normal distribution)

4-1 χ^2 分布 (Chi–square distribution)

■ 定義

正規分布 $n(x;\ \mu_i,\sigma_i)$ に従う独立な ν 個の標本を x_i (i=1, 2,⋯ ν) とすると,

$$\chi^2 = \sum_{i=1}^{\nu} \left(\frac{x_i-\mu_i}{\sigma_i}\right)^2 \quad \text{は}$$

$$f(\chi^2,\nu) = \frac{1}{2^{\nu/2}\,\Gamma(\nu/2)}\chi^{2(\nu/2-1)}\exp\left(-\frac{\chi^2}{2}\right) \tag{4-1}$$

を持つ自由度 (degree of freedom) ν の χ^2 分布に従う. ただし Γ はガンマ関数 (gamma function)

$$\Gamma(x) = \int_0^{\infty} e^{-u}u^{x-1}du,\ 0<x<\infty \tag{4-2}$$

■ 導出

$$\chi^2 = \sum_{i=1}^{\nu} \left(\frac{x_i-\mu_i}{\sigma_i}\right)^2 = \Sigma z_i^{\,2} \text{ とする.} \tag{4-3}$$

z_i は正規分布 $n(z_i;\ 0, 1)$ に従う.

$u_i = \Sigma z_i^2$ とする. その分布関数は

$$f(u_i) = \frac{1}{\sqrt{2\pi\,u_i}}\exp\left(-\frac{u_i}{2}\right) \qquad [\leftarrow\text{(1-22)参照}] \tag{4-4}$$

この特性関数は

$$\phi_i(t) = \int_0^{\infty} \frac{1}{\sqrt{2\pi\,u_i}}\exp\left(-\frac{u_i}{2}\right)\exp(itu_i)\,du_i$$

$$= (1-2it)^{-1/2} \tag{4-5}$$

u_i は互いに独立であることから

$$\phi(t) = \prod_{i=1}^{\nu}\phi_i(t) = (1-2it)^{-\nu/2} \tag{4-6}$$

よって

$$f(\chi^2,\nu) = \frac{1}{2\pi}\int_0^{\infty}(1-2it)^{-\nu/2}\exp(-i\chi^2 t)\,dt \tag{4-7}$$

→ (4-1) になる.

MGF

$$M(t) = (1-2t)^{-\nu/2} \tag{4-8}$$

$\rightarrow \mu = \nu,\ \sigma^2 = 2\nu,\ \mu_3 = 8\nu,\ \mu_4 = 12\nu(\nu+4)$

よって，skewness $\beta_1 = 8/\nu$, kurtosis $\beta_2 = 3(1+4/\nu)$

■ χ^2 分布の性質

・加法性 (additivity)

χ_1^2 が自由度 ν_1 の χ^2 分布，χ_2^2 が自由度 ν_2 の χ^2 分布に従うとき，$\chi_1^2+\chi_2^2$ は自由度 $\nu_1+\nu_2$ の χ^2 分布に従う．

証明

特性関数の性質により

$$\begin{aligned}\phi_{\chi_1^2+\chi_2^2}(t) = \phi_{\chi_1^2}(t)\ \phi_{\chi_2^2}(t) &= (1-2it)^{-\nu_1/2}\ (1-2it)^{-\nu_2/2} \\ &= (1-2it)^{-(\nu_1+/\nu_2)/2}\end{aligned} \tag{4-9}$$

・χ^2 分布は $\nu\rightarrow\infty$ で正規分布 $n(\chi^2, \nu, 2\nu)$になる．

証明

$$y = \frac{\chi^2-\mu}{\sigma} = \frac{\chi^2-\nu}{\sqrt{2\nu}}$$ とすると，特性関数は

$$\phi_y(t) = \exp(-\frac{i\nu t}{\sqrt{2\nu}})(1-\frac{2it}{\sqrt{2\nu}})^{-\nu/2} \tag{4-10}$$

すなわち

$$\log\phi_y(t) = -\frac{i\nu t}{\sqrt{2\nu}} - \frac{\nu}{2}\log\ (1-\frac{2it}{\sqrt{2\nu}}) \tag{4-11}$$

$\nu\rightarrow\infty$ で，右辺の log を展開すると

$$\log\phi_y(t) = -\frac{i\nu t}{\sqrt{2\nu}} - \frac{\nu}{2}(-\frac{2it}{\sqrt{2\nu}} + \frac{1}{2}\ \frac{4t^2}{2\nu} + \cdots) = -\frac{t^2}{2}$$

$$\therefore\ \phi_y(t) = \exp(-\frac{t^2}{2}) \tag{4-12}$$

・$x_1, x_2, \cdots x_\nu$ が正規分布 $n(x; 0, 1)$の標本なら，

$$u = \sum_{i=1}^{\nu}(x_i-\bar{x})^2 \quad \text{ただし}\ \bar{x} = \frac{1}{\nu}\Sigma x_i$$

は自由度$(\nu-1)$の χ^2 分布に従う．

導出

$$\begin{aligned}&u_1 = (x_1-x_2)/\sqrt{2},\ u_2 = (x_1+x_2-2x_3)/\sqrt{6},\ \cdots \\ &\quad u_{\nu-1} = (x_1+\cdots-x_{\nu-1}-(\nu-1)x_\nu)/\sqrt{\nu(\nu-1)}, \\ &\quad u_\nu = (x_1+\cdots+x_\nu)/\sqrt{\nu}\end{aligned} \tag{4-13}$$

と置くと，

u_i も $n(u_i;0,1)$に従う．よって，

$$u = \sum_{i=1}^{\nu} (x_i - \bar{x})^2 = \sum_{i=1}^{\nu} x_i^2 - \nu\bar{x}^2 = \sum_{i=1}^{\nu} u_i^2 - u_\nu^2 = \sum_{i=1}^{\nu-1} u_i^2 \tag{4-14}$$

は自由度$(\nu-1)$の χ^2 分布.

母集団の分散が σ^2 なら，$Z = \dfrac{1}{\sigma^2}\sum_{i=1}^{\nu}(x_i - \bar{x})^2$ は自由度$(\nu-1)$の χ^2 分布.

・標本分散 (不偏分散) の性質

$$s^2 = \frac{1}{\nu-1}\Sigma(x_i - \bar{x})^2 = \frac{\sigma^2 Z}{\nu-1} \tag{4-15}$$

よって，$(\nu-1)s^2/\sigma^2$ は$\bar{x}$によらずに自由度$(\nu-1)$の χ^2 分布.

4-2 *t* 分布 (Student's *t* distribution)

■ 定義

u が $\mu = 0$, $\sigma^2 = 1$ の正規分布，ω が自由度 ν の χ^2 分布に従い，u と ω が独立であるとき，

$t = \dfrac{u}{\sqrt{\omega/\nu}}$ は自由度 ν の t 分布に従う.

$$f(t;\nu) = \frac{\Gamma[(\nu+1)/2]}{\sqrt{\nu\pi}\,\Gamma(\nu/2)}\left[1 + \frac{t^2}{\nu}\right]^{-(\nu+1)/2} \qquad (-\infty < t < \infty) \tag{4-17}$$

■ 導出

u,ω の joint density function は

$$f(u,\omega;\nu) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{u^2}{2}\right) \times \frac{1}{2^{\nu/2}\Gamma(\nu/2)}\omega^{\nu/2-1}e^{-\omega/2}$$

$u = t(\omega/\nu)^{1/2}$ と置けば

$$= \frac{1}{\sqrt{2\pi\nu}\,2^{\nu/2}\Gamma(\nu/2)}\exp\left(-\frac{t^2}{\omega/2\nu}\right)\omega^{\nu/2-1/2}e^{-\omega/2} \tag{4-18}$$

t の分布関数を求めるには，これを ω で積分する.

$$f(t,\nu) = \int_0^\infty f(t,\omega;\nu)\,d\omega \tag{4-19}$$

Γ分布の式 (4-2) を使うと (4-17) が出る.

■ *t* 分布の性質

・x_i $(i=1,\cdots,n)$ が正規分布 $n(\mu,\sigma^2)$ に従う母集団から取った標本であるとき，

$u=\dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}}$は正規分布 $n(0,1)$, $\omega=\dfrac{(n-1)s^2}{\sigma^2}$は自由度$(n-1)$の χ^2 分布に従う

$\rightarrow$ $t=\dfrac{\sqrt{n}}{s}(\bar{x}-\mu)$ は自由度$(n-1)$の t 分布に従う．　(4-20)

・$\nu\rightarrow\infty$ で $f(t,\nu)\rightarrow n(0,1)$

導出：Stirling の式 $\Gamma(\nu+1)\rightarrow\sqrt{2\pi}\,\nu^{(\nu+1)/2}e^{-\nu}$ による．式変形省略．

・2 つの標本平均の差の分布
正規分布をする 2 つの母集団 $n(x_1,\mu_1,\sigma^2)$, $n(x_2,\mu_2,\sigma^2)$ から n_1 個, n_2 個の標本を取る．それぞれの平均

$$\bar{x}_1=\frac{1}{n_1}\sum_{i=1}^{n_1}x_{1i},\quad \bar{x}_2=\frac{1}{n_2}\sum_{i=1}^{n_2}x_{2i}, \tag{4-21}$$

について，

$$t=\frac{[(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)]}{[Sp^2\,(\frac{1}{n_1}+\frac{1}{n_2})]^{1/2}} \tag{4-22}$$

は自由度 $\nu=n_1+n_2-2$ の t 分布に従う．ただし

$$Sp^2=\frac{\sum_{j=1}^{2}\sum_{i=1}^{n_j}(x_{ji}-\bar{x}_j)^2}{(n_1+n_2-2)} \tag{4-23}$$

導出

$Sp^2=\dfrac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1+n_2-2)}$ と書ける．

χ^2 分布の additivity から，
$\omega=Sp^2(n_1+n_2-2)/\sigma^2$ は自由度 $\nu=n_1+n_2-2$ の χ^2 分布に従う．

また，$\bar{x}=\bar{x}_1-\bar{x}_2$ は平均 $\mu=\mu_1-\mu_2$, 分散 $\sigma_d^2=\sigma^2/n_1+\sigma^2/n_2$ の正規分布をなす [(3-19), (1-19)] ので，

$$u=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{[\sigma^2\,(\frac{1}{n_1}+\frac{1}{n_2})]^{1/2}}=\frac{\bar{x}-\mu}{\sigma_d} \tag{4-24}$$

は平均 0, 分散 1 の正規分布．$\bar{x}_i$ と s_i^2 は独立なので，u と ω も独立で，
$t=u/[\omega/(n_1+n_2-2)]^{1/2}$ は自由度 n_1+n_2-2 の t 分布をなす．

・t 分布は対称

$$P[t<-t_\alpha(\nu)] = P[t>t_\alpha(\nu)] = \alpha \tag{4-25}$$

・t 分布の数値は t 分布表
近似式 (Wallace, 1959): 正規分布による確率点 x_α に対応する t 分布の確率点 t_α

$$t_\alpha^2(\nu) = \nu \exp\{\frac{x_\alpha^2[1+2/(1+8\nu)]^2}{\nu} - 1\} \tag{4-26}$$

もっと高精度の式もある．例：山内 (1968)

4-3 F 分布 (F distribution)

■ 定義

2 つの独立変数 χ_i^2 $(i=1, 2)$ が自由度 ν_i の χ^2 分布なら,

$F = \dfrac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2}$ は

$$f(F;\nu_1,\nu_2) = \frac{\Gamma[(\nu_1+\nu_2)/2]}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} (\nu_1/\nu_2)^{\nu_1/2} \frac{F^{(\nu_1-2)/2}}{[1+(\nu_1/\nu_2)F]^{(\nu_1+\nu_2)/2}} \tag{4-27}$$

の F 分布をなす.

平均 $\mu = \nu_2/(\nu_2-2)$ (4-28)

分散 $\mathrm{var}(F) = 2\nu_2(\nu_1+\nu_2-2)/\nu_1(\nu_1-2)^2(\nu_2-4)$ (4-29)

2 つの分散，または 2 つ以上の平均を比べるときに使う.

■ 導出

$u = \chi_1^2$, $w = \chi_2^2$ と置く．joint density function は

$$g(u,w) = \frac{u^{(\nu_1-2)/2} w^{(\nu_2-2)/2}}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)\, 2^{(\nu_1+\nu_2)/2}} \exp[-\frac{1}{2}(u+w)] \tag{4-30}$$

$u = (\nu_1/\nu_2)\, wF$ と置けば
F,w の joint density function は

$$f(F,w) = \frac{w^{(\nu_2-2)/2}}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)\, 2^{(\nu_1+\nu_2)/2}} (\frac{\nu_1}{\nu_2} w)(\frac{\nu_1 w}{\nu_2}F)^{(\nu_1-2)/2} \exp[-\frac{w}{2}(1+\frac{\nu_1}{\nu_2})F] \tag{4-31}$$

これを w について積分する

$$f(F;\nu_1,\nu_2) = \frac{F^{(\nu_1-2)/2}}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)\, 2^{(\nu_1+\nu_2)/2}} (\frac{\nu_1}{\nu_2})^{\nu_1/2} I(F,\nu_1,\nu_2) \tag{4-32}$$

$$I(F;\nu_1,\nu_2) = \int_0^\infty w^{(\nu_1+\nu_2-2)/2} \exp[-\frac{w}{2}(1+\frac{\nu_1}{\nu_2})F]\, dw$$

$$= \frac{\Gamma[(\nu_1+\nu_2)/2]\ 2^{(\nu_1+\nu_2)/2}}{(1+\nu_1 F/\nu_2)^{(\nu_1+\nu_2)/2}} \tag{4-33}$$

■ F 分布の性質

・パーセント点

$$P[F \geqq F_\alpha] = \alpha = \int_{F_\alpha}^{\infty} f(F;\nu_1,\nu_2)\,dF \tag{4-34}$$

について

$$F_{1-\alpha}(\nu_1,\nu_2) = [F_\alpha(\nu_2,\nu_1)]^{-1} \tag{4-35}$$

・χ^2 分布との関係

$\nu \to \infty$ で，$|(\chi^2/\nu)-1| \to 0$, よって $F(\nu_1,\infty) = \chi_1^2/\nu_1$ (4-36)

すなわち χ^2 分布は F 分布の特殊な場合.

・t 分布との関係

$\nu_1 = 1$ の場合，$\chi_1^2/\nu_1 = u^2$ とすると u は正規分布 $n(0,1)$に従う．よって

$$F(1,\nu_2) = \frac{u^2}{\chi_2^2/\nu_2} \tag{4-37}$$

$t = u/(\chi_2^2/\nu_2)$ は自由度 ν_2 の t 分布をするので，$F(1,\nu_2) = t^2(\nu_2)$

Ⅴ 主成分分析 (Principal component analysis, PCA)

5-1 主成分分析の考え方 (Basic idea)

・多次元のデータ

例：x_i; 地点 i の気温

例：x_i; 科目 i の成績

これの変動の「主たる特徴」を見出す.

すなわち,「主たる特徴」を表すような, x_i の 1 次結合 y

$$y = \sum_{i=1} a_i x_i \tag{5-1}$$

を求める.

言い替えると,「主たる特徴」が表現されるように, 係数 a_i を決める.

■ 2 種類の考え方

(1) y_j の分散が最大になる方向へ, 新しい座標軸を取る.

「分散が最大になる」とは, データの変動 (データ同士の違い) が最も強く表現されることを意味する. そうなるように, 係数 a_i を決める.

次に,「上記の座標軸と直交し, かつ分散が最大になる方向へ 2 つめの座標軸を取る」. 最初の係数を a_{i1}, 今回のを a_{i2}, …

このようにして, 互いに直交する座標軸を順次求めていく.

k 番目の座標は

$$y_k = \sum_{i=1} a_{ik} x_i \tag{5-2}$$

(2) 各成分間の相関 (共分散) が 0 になるように, 座標軸を取る.

これは相関行列 (共分散行列) を対角化し, 座標を直交変換することに他ならない.

歴史的に見ると, 主成分分析の初期の概念は(1)だった. この場合, y_j の分散を最大にする係数 a_i を求めることを, 主成分の「抽出」と称した. 当時の計算機事情では, すべての主成分を一気に求める ((2) の方法で) のは困難であり, 苦労して順番に成分を求めていったという事情が反映されている.

しかし, (1)と(2)は数学的には同じである.

現在では, 単純で分かりやすい(2)の考え方が採用されることが多い.

それとともに, 定式化のやり方も変わってきている (規格化のしかた等. 後述).

5-2 PCA の定式化 (Formulation)

■ 古典的な PCA の定式化 (上記(2)による. 以下では, $x \to z$, $y \to f$ と表記する)

地点 $i = 1 \sim N$, 時刻 $j = 1 \sim J$ として, $\{z_{ji}\}$ を $N \times J$ のデータがあるとする.

これの 1 次結合

$$z_{ji} = \sum_{k=1}^{N} f_{jk} a_{ik} \tag{5-3}$$

ただし z の時間平均は 0 とする (＝時間平均を差し引いてある). $\sum_{j=1} z_{ji} = 0$

f は正規直交 (orthonormal) 条件を満たす.

$$\frac{1}{J}\Sigma\ f_{jk}\, f_{j\ell} = 1 \text{ (if } k = \ell\text{)} \qquad = 0 \text{ (if } k \neq \ell\text{)} \tag{5-4}$$

上式の行列表現

$$Z = FA' \tag{5-5}$$

$$\frac{1}{J}F'F = I \quad \text{(単位行列)} \tag{5-6}$$

付記：「i が地点，j が時刻」である必要はない．例：i が科目，j が個人，z がテストの得点

■ A の求め方

z の共分散行列 (covariance matrix)

$$C = \frac{1}{J}Z'Z = \frac{1}{J}AF'FA' = AA' \tag{5-7}$$

これは C の対角化に相当する.

$$D = T'CT \tag{5-8}$$

ただし,

$$D = \begin{pmatrix} \lambda_1^2 & & 0 \\ & \ddots & \\ 0 & & \lambda_N^2 \end{pmatrix} \quad : \text{対角行列; } \lambda^2 \text{ は固有値} \tag{5-9}$$

T は固有ベクトルから成る直交行列.

$$C = TDT' = (TD^{1/2})\,(TD^{1/2})' \text{ となるので，} A = TD^{1/2} \text{ と置ける．} \tag{5-10}$$

A の直交性

$$A'A = D^{1/2}T'TD^{1/2} = D \tag{5-11}$$

すなわち，F だけでなく A も直交する．言い替えると，主成分分析の結果は，時間的にも空間的にも直交する.

F は無次元，A は有次元 (Z と同じ次元) であることに注意.

F をスコア (score)，A をパタン (pattern) あるいは負荷 (loading) と言う.

■ 寄与率

分散全体に対する各成分の寄与

$$\text{全体の分散 } V = \frac{1}{J}\sum_{i,j} z_{ji}^2 = \mathrm{tr}(C) = \sum_{k,i} a_{ik}^2 = \sum_k \lambda_k^2$$

$$\text{第 } k \text{ 成分による分散 } V_k = \frac{1}{J}\sum_{i,j,k} (f_{jk}\, a_{ik})^2 = \sum_i a_{ik}^2 = \lambda_k^2$$

寄与率 $c_k = \sum_i a_{ik}^2 \,/\, V$ (5-12)

$= \lambda_k^2 \,/\, V$ (5-13)

■ F の求め方

式(5-5)から,

$ZA = FA'A = FD$, よって $F = ZAD^{-1}$ (5-14)

■ 補足 1: 規格化の流儀

上記では，時間項 F を規格化した.

別の流儀

① A も規格化する.

$A = TD^{1/2}$ ではなく，$A^* = T$ とする.

この場合，$Z = FD^{1/2}A^{*\prime} = FD^{1/2}T'$ と表される.

$A^* = T$ は「固有ベクトル」(eigenvector) と呼ばれる.

初期の主成分分析はこの流儀のものが多い.

② A を規格化し，F は規格化しない.

前項の式で $F^* = FD^{1/2}$ とすると，$Z = F^*T' = F^*A^{*\prime}$

この場合にはスコア F*が有次元，パタン T が無次元となり，直交条件は

$$T'T = I, \quad \frac{1}{J}F^{*\prime}F^* = D \tag{5-15}$$

PCA を使った論文を書くときは，どの流儀によるのかを明示する必要がある.

・補足 2：相関行列を使う方法

z_{ji} を規格化しておく: $z_{ji}^* = z_{ji} / \sigma_i$

i が次元の異なる変数を含むときは，この操作が必要.

例:

$i=1$ は日平均気温，$i=2$ は日照時間，…

次元が同じでも，地点によって変動幅 (分散) に大きな差がある場合

規格化すべきかどうかは解析の目的次第.

・補足 3：主成分分析の解釈に関する注意事項

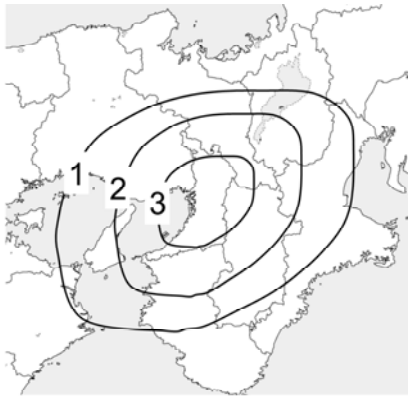
降水量の主成分分析を行ったところ，ある成分が右のパタンになったとしよう.

この結果から「大阪付近は降水量が多い」と解釈していいか？

答. 2 つの理由でダメ.

(1) 主成分分析は変動のパタンを求めるものであり，ナマのパタンを求めるものではない.

(2) これは，変動を分解したものの 1 成分に過ぎない．真の変動は，多数のパタンの和である.



・補足 4：$N > J$ の場合には，i と j を入れ替えた形で主成分分析を行い，式 (5-15) の要領で戻す方法がある (dual formalism).

5-3 回転 PCA (Rotated PCA)

■ 普通の PCA (上記の方法＝非回転 PCA, unrotated PCA) の問題点：

F (時間変化の係数) だけでなく A (空間パタン) も直交する．このことが，A の特徴を規定する．

多くの場合，

第 1 成分：全域同符号

第 2 成分：正域と負域のシーソー

高い成分になるほど，複雑なパタンになっていく

これは，対象領域の全体的な変動を抽出するにはいいが，地域ごとの変動を見るには必ずしも適しない．

■ 回転 PCA とは：

A の直交条件 (5-11) に代え，地域ごとの特徴が現れやすい条件を使う．その条件を満たすように，空間座標を回転 (＝直交変換) する．

A に任意の直交変換を施しても，スコアの直交性は維持される．

証明：

K 個の主成分を表す A ($N \times K$ 次元) に対し，直交変換 T ($K \times K$ 次元) を施すとする．

直交変換の性質上，TT′ = T′T = I (単位行列．この T は前ページの T とは別).

$$B = AT,\ G = FT \text{ (従って, } A=BT',\ F=GT') \tag{5-16}$$

とすれば

$$Z = FA' = GTT'B' = GB' \tag{5-17}$$

$$\frac{1}{J}G'G = \frac{1}{J}T'F'FT = T'T = I \tag{5-18}$$

となり，(5-5), (5-6)と同じ形になる．ただし，A の直交性 (5-11) は外れる．

$$B'B = T'A'AT = T'DT \tag{5-19}$$

D は対角行列だが，T′DT は対角行列にはならない．

寄与率は･･･(5-13) はダメだが (5-12) は成り立つ．

では，直交条件 (5-11) に代えて，どんな条件をつけようか？

■ varimax 条件：

主成分負荷 a_{ik} の 2 乗の分散を最大にする．

これは，主成分負荷の分布ができるだけ局所化されることを意味する (単純構造＝simple structure).

第 k 成分の負荷の 2 乗の分散は

$$V_k = \frac{1}{N}\sum_{i=1}^{N}(b_{ik}^2 - \frac{1}{N}\sum_{h=1}^{N} b_{hk}^2)^2 \tag{5-20}$$

これをすべての k について合計すると

$$V = \sum_{k=1}^{K} [\frac{1}{N} \sum_{i=1}^{N} (b_{ik}^2 - \frac{1}{N} \sum_{h=1}^{N} b_{hk}^2)^2] \rightarrow \max. \quad (5\text{-}21)$$

解法については文献参照．芝　祐順「因子分析法」(1979, 東京大学出版会) など．

・raw varimax と normal varimax

上記は raw varimax method と呼ばれる．

b_{ik}の代わりに, これを規準化した b_{ik}/s_i を使うものを normal varimax method と言う．

ただし

$$s_i^2 = \sum_{k=1}^{K} b_{ik}^2 \quad (5\text{-}22)$$

なお，PCA に相関行列を使えば，raw varimax と normal varimax とは同一である．

■ 重みつき varimax

地点 i の重みを v_i として,

$$V = \sum_{k=1}^{K} [\sum_{i=1}^{N} v_i (b_{ik}^2 - \frac{\sum_{h=1}^{N} v_h b_{hk}^2}{\sum_{h=1}^{N} v_h})^2] \quad (5\text{-}23)$$

■ varimax method の一般化

$$V = \sum_{k=1}^{K} [\frac{1}{N} \sum_{i=1}^{N} (b_{ik}^2 - \frac{w}{N} \sum_{h=1}^{N} b_{hk}^2)^2] \rightarrow \max. \quad (5\text{-}24)$$

を orthomax criterion と言う．varimax method は w=1，非回転 PCA は $w = -\infty$ の場合に該当する．

その他，w の値によって次のように呼ばれる．

w=0: quartimax method

w=1/2: biquartimax method

w=K/2: equamax method

他にもいくつかの回転方法があるが (芝, 1979), いくつかの気象データに適用してみた印象では，varimax method による結果が最も自然であるように思えた．

■ 回転する成分数 K の決め方

K が十分大きければ，varimax　rotation で得られる主要なパタンは K に依存しない．

従って，K を十分大きく与える (30 とか) のがお勧め．

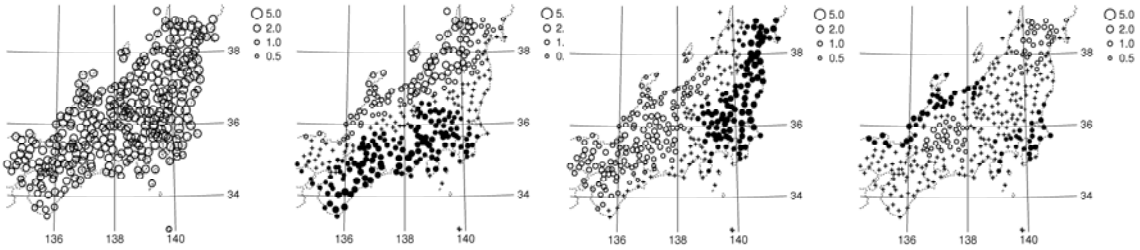
■ 補足

非回転 PCA を「主因子法」principal axis method, それの結果を主因子解 (principal axis solution) と言うことがある.

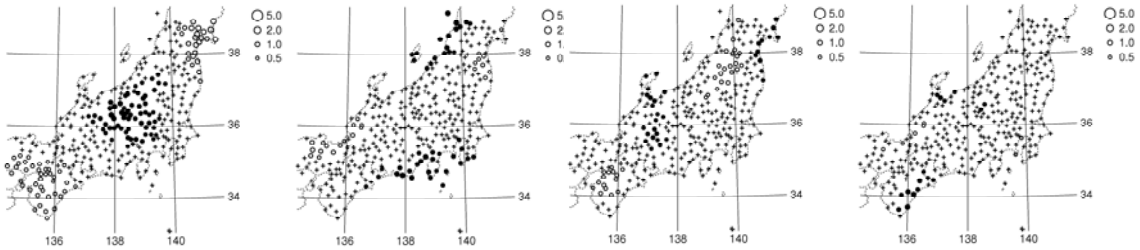
■ 解析例:

本州中部の気温 (daily 12～15JST, July and August, 1979 ～ 2014)

・非回転解 (unrotated analysis based on covariance matrix)

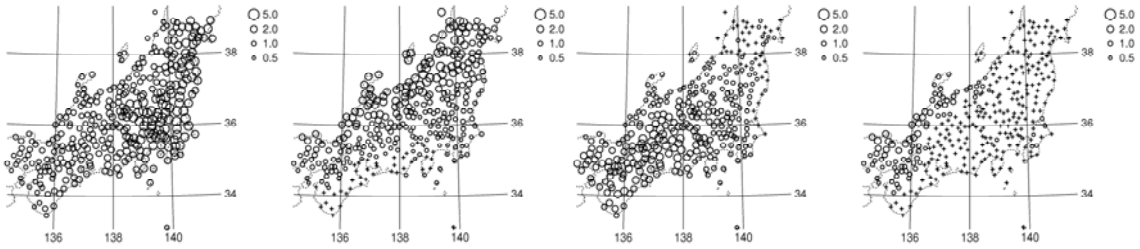


k=1 (68.0%) *k*=2 (8.9%) *k*=3 (6.9%) *k*=4 (2.3%)

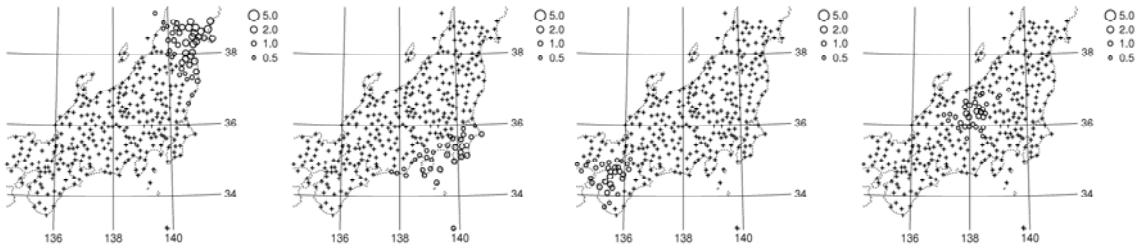


k=5 (2.1%) *k*=6 (1.2%) *k*=7 (1.0%) *k*=8 (0.7%)

・varimax 回転解 (rotated analysis using raw varimax, *K*=50)



k=1 (40.6%) *k*=2 (21.6%) *k*=3 (17.9%) *k*=4 (3.2%)



k=5 (2.6%) *k*=6 (0.9%) *k*=7 (0.8%) *k*=8 (0.8%)

5-4 主成分分析，経験的直交関数展開，因子分析の違い (Differences of PCA, empirical orthogonal functions, and factor analysis)

これらは，計算手段が貧弱だった時代に，別々に発展してきたという経緯がある．
そのため概念的には異同があるが，数学的な本質は同じだと考えられる．

■ 主成分分析と因子分析の違い

因子分析では「独自因子 (specific factor, unique factor)」を考える．
主成分分析の式は

$$Z = FA'$$

すなわち

$$z_{ji} = \sum_{k=1}^{N} f_{jk}\, a_{ik}$$

$$= \sum_{k=1}^{K} f_{jk}\, a_{ik} + \sum_{k=K+1}^{N} f_{jk}\, a_{ik} \qquad (5\text{-}25)$$

因子分析の式は

$$z_{ji} = \sum_{k=1}^{K} f_{jk}\, a_{ik} + d_{ji} \qquad (5\text{-}26)$$

↑
独自因子

もし式(5-25) の右辺第 2 項 (第 K+1 成分以下の項) を独自因子と見なせば，主成分分析と因子分析は数学的に同一となる．多くの実用場面では，そうなっている (多分).

(←「考え方が違うのだから，あくまで別の方法だ」という主張もあるが…)

■ 類似の方法

特異値分解
独立成分分析
正準相関分析

VI 極値統計 (Extreme value analysis)

6-1 極値統計の基本概念 (Basic idea)

■ 極値統計とは

極端な事象 (extreme events; 大雨，強風，高温・低温など) が起きる頻度 (frequency) や強度 (intensity) を，長期間のデータから推定する．

・極値統計の社会的意義 (application of extreme value analysis) :

防災対策 (disaster countermeasures) の目安．建造物の設計雨量 (design rainfall)，設計風速 (design wind speed) など．

■ 再現期間とは何か

・再現期間 (return period): ある事象の発現確率を表す数値．再来期間とも言う．

「再現期間 100 年」とは：その事象が 100 年に 1 回の確率で起きる．言い替えると，1 年間の発現確率は 100 分の 1.

・再現期待値 (return value) : 再現期間に対応する値.

「100 年再現降水量」「50 年再現風速」などの言い方もある．

・再現期間 T 年の事象が，t 年間に起きる確率は？

1 年間にその事象が起きない確率は $(1-1/T)$.

t 年間にその事象が起きない確率は $(1-1/T)^t$.

t 年間にその事象が起きる確率は $1-(1-1/T)^t$.

$t=T$ の場合，$1-(1-1/T)^T \fallingdotseq 1-1/e = 0.63$. (6-1)

$\because \lim_{x\to\infty}(1-1/x)^x = 1/e$

■ 再現期間についてのよくある誤解 (misunderstanding about return period)

・「再現期間＝周期」という誤解 (confusion of return period and periodicity)

正しくは，再現期間は確率 (probability) に過ぎない．従って，再現期間 100 年の大雨が 2 年続いて起きることもあり得る．宝クジに 2 回続けて当たるようなもの．

(巨大地震は実際に周期性があるため，余計に誤解されやすい)

・「再現期間 100 年」→「100 年間でたったの 1 回」→まずあり得ないこと･･･という誤解

実際には，再現期間 100 年の事象の発生確率は，50 代でガンになる確率よりも高い.

6-2 極値統計の数学的基礎 (Mathematical basis)

■ 極値統計の基本原理

・同じ確率分布に従う互いに独立な (identically indipendently distributed = iid) 十分多数のデータの極値は,

その確率分布が何であっても,

一般化極値分布 (generalized extreme value distribution = GEV) になる (収束する).

・日々の気象観測値 (daily observation value) を 1 年分集めれば，これらの条件を満たすのではないか･･･

→ それなら年極値 (annual extreme) に GEV を適用し，再現期間などを推定することができる．

■ GEV の定式

・GEV の累積分布関数 (CDF) ＝非超過確率 (non-exceedance probability)

$$F(x) = \exp[-\{1 - \frac{\kappa(x-\beta)}{\alpha}\}^{1/\kappa}], \ (\kappa \neq 0) \tag{6-2}$$

$$F(x) = \exp\{-\exp(-\frac{x-\beta}{\alpha})\}, \ (\kappa = 0) \tag{6-3}$$

β: location parameter (位置パラメーター．ただし空間的位置という意味ではない)
α: scale parameter (尺度パラメーター)
κ: shape parameter (形状パラメーター).

$\kappa > 0$: 極端な値は出にくい.
$\kappa < 0$: 低頻度で極端な値が出やすい．κの絶対値が大きいほど，その傾向は強まる．降水の場合，時間スケールが長いほど κは負方向へシフトする．

・値 x の再現期間

$$T(x) = 1/(1-F(x)) \tag{6-4}$$

再現期待値: $F(x)$の逆関数

$$x(F) = \beta + \frac{\alpha}{\kappa}\{1 - (-\log F)^{\kappa}\}, \quad (\kappa \neq 0) \tag{6-5}$$

$$x(F) = \beta - \alpha \log(-\log F), \quad (\kappa = 0) \tag{6-6}$$

■ 補足

・「パラメーター」を「母数」とも言う
・κ の符号を上式と反対にする流儀もあるので要注意 (その場合は ξ と書かれることが多い).
・数学的には
$\kappa = 0$: x に上限も下限もない場合.
$\kappa > 0$: x に上限がある場合.
$\kappa < 0$: x に下限がある場合.

上・下限値は(6-2)より $x = \beta + \frac{\alpha}{\kappa}$ (6-7)

しかし，この上限は数学上のものであり，実用場面ではデータに合うように κ を決めるのが一般的.
日本の年最大日降水量の場合，α=30 ～ 40mm, β=100 ～ 150mm ぐらい
→ κ = 0.05 なら上限は 2000 ～ 3000mm ≫現実的上限

・歴史的には，$\kappa = 0$, $\kappa > 0$, $\kappa < 0$ の 3 者が別々に導かれた.
$\kappa = 0$: Gumbel distribution
$\kappa > 0$: Weibull distribution

$\kappa < 0$: Fréchet distribution

その後，GEV として統合された．

6-3 極値統計の手法 (Methods)

分布 parameters の求め方 (estimation of parameters)

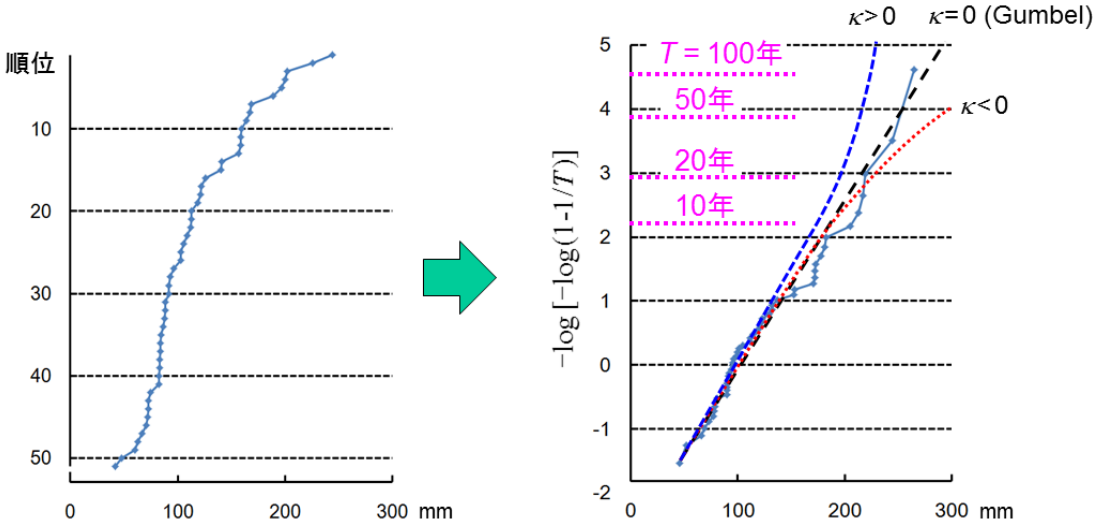
■ graphical method

計算資源が貧弱だった時代は，計算図表を使った図式的な方法が使われた．Gumbel 分布を前提とする．

二重指数確率紙 (double exponential probability paper)：横軸に x を取り，縦軸に式 (6-6) 右辺の $-\log(-\log F)$すなわち $-\log[-\log(1-1/T)]$を取ったもの．Gumbel 分布は直線で表現される．

$T \to$大なら $\log(1-1/T) \to -1/T$, $\therefore -\log[-\log(1-1/T)] \to -\log(1/T) = \log T$.

すなわち，二重指数確率紙の縦軸はほぼ再現期間の対数になっている．



・解析手順

(1) 二重指数確率紙にデータを描き込んでいく．その際，データ年数を N として，1 位のデータの再現期間を $T = N/0.5$, 2 位のデータは $T = N/1.5$, ･･･, i 位のデータは $T = N/(i-0.5)$と見なす．

(2) 描きこんだデータに直線を当てはめる．その傾きが α, x 切片が β を与える．

・補足：

(1) における T の与え方 (plotting position) は，上記以外にもいろいろ提案されている．

$T = N/(i-0.5)$: Hazen plot (上記の方法)

$T = (N+0.2)/(i-0.4)$: Cunnane plot

$T = N/i$: California plot

$T = (N+1)/i$: Weibull plot

「数学的には Weibull plot が正しい (only one correct plotting position)」(Makkonen, 2006, JAMC)

「plotting position よりも，パラメーターを求めるほうが効率的 (Fitting a Gumbel

distribution can be done much more efficiently by estimating its parameters)」(de Haan, 2007, JAMC)

■ 積率法 (moment method)

・例として Gumbel 分布の場合，mean と variance は

$$\mu = \beta - \gamma\alpha \tag{6-8}$$

$$\sigma^2 = \pi^2\alpha^2/6 \tag{6-9}$$

で与えられる．ただし γ は Eiler's constant = 0.5772･･･

そこで，N 年間の年極値データの平均値と分散

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{6-10}$$

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \overline{x})^2 \tag{6-11}$$

を計算し，これらを上式の μ, σ^2 に当てはめて α と β を求める．

GEV の場合はパラメーターが 3 つあるので，mean，variance に加えて 3 次の moment (skewness) の式を使う．

L-moments の普及につれ，積率法は使われなくなった．

■ L-moment 法 (L-moment method)

・積率法と同様だが，variance や skewness の代わりに L-moments を使う．

Gumbel 分布の場合，

$$\alpha = \frac{\lambda_2}{\log 2}, \tag{6-12}$$

$$\beta = \lambda_1 - \alpha\gamma, \tag{6-13}$$

L-moments は，極端な外れ値 (outlier) に影響されにくく，積率法よりも優れているとされる．

Hosking は，L-moments を計算する Fortran プログラムを公開している:
http://lib.stat.cmu.edu/general/lmoments

■ 最尤法 (maximum likelihood method; MLE)

・極値分布関数の PDF を $F(x; \alpha, \beta, \kappa)$ のように表記すると，データ $x_1, x_2, \cdots x_N$ についての尤度関数は

$$L(x_1, x_2, \cdots x_N; \alpha, \beta, \kappa) = \prod_{i=1} f(x_i; \alpha, \beta, \kappa) \tag{6-14}$$

$L(\theta)$を最大にする α, β, κ を求めるため，

$$\frac{\partial}{\partial\alpha}\log L(\theta) = 0, \quad \frac{\partial}{\partial\beta}\log L(\theta) = 0, \frac{\partial}{\partial\kappa}\log L(\theta) = 0, \tag{6-15}$$

$\log L(\theta)$の微分にデータ $x_1, x_2, \cdots x_N$ の値を入れて上記の方程式を数値的に解くことにより，α, β, κの最尤推定値 (ML estimate)を求めることができる．

L-moments のプログラムが提供されていない分布関数 (例えば平方根指数型最大値分布＝ square-root-exponential type maximum distribution; SQRT-ET) を使うときは，最尤法が適している．

補足: 数値解法として Newton 法を使う場合，$\log L(\theta)$の 2 階微分が必要になる．計算面倒．

6-4 極値統計の精度 (Confidence of extreme value analysis)

■ 極値統計の誤差要因

(1) データの確率変動 (random variability of data)

(2) 理論の前提が不成立 (inconsistency of the theory and the data)

iid 条件が不成立 (failure of the iid condition)

気候の非定常性 (non-steadiness of the climate)

■ 適合度 (degree of fitness) による関数選択

日々の気象データは，実際には iid 条件を満たさない (程遠い).

従って，年極値が GEV に従う保証はない.

そこで，GEV を含む複数の分布関数を用意し，その中からデータに適合するものを選ぶという方法が取られることがある.

・分布関数として使われるもの

Gumbel 分布，GEV

SQRT-ET

対数ピアソンⅢ型分布 (log Pearson type Ⅲ distribution; LP3)

対数正規分布 (generalized normal distribution; GNO)

等々．詳細は文献参照.

■ 適合度基準 (goodness-of-fit criterion)

日本では SLSC がよく使われる.

・SLSC (standard least-squares criterion) とは:

データを確率紙にプロットして関数を当てはめたときの，データ値と関数値の差 (それぞれ標準化した値) を 2 乗平均したもの.

$$\mathrm{SLSC} = \frac{\sqrt{\sum_{i=1}^{N} (s_i - s_i^*)^2}}{N\,|s_{0.99} - s_{0.01}|} \tag{6-16}$$

ただし，大きい方から i 番目のデータを x_i として

$s_i = -\log\{-\log F(x_i)\}$: データ値に対応する Gumbel 分布の関数値

$s_i^* = -\log[-\log\{(i-0.5)/N\}]$: 順位に相当する関数値

$s_{0.99}$, $s_{0.01}$ はそれぞれ関数の 99%値, 1%値に対応する s_i 値

すなわち $s_{0.99} = -\log(-\log 0.99)$, $s_{0.01} = -\log(-\log 0.01)$

・SLSC が小さいほど，適合性が高い．では，SLSC がいくつ以下ならいいか？
高棹ほか (1986)，宝・高棹 (1988): 「SLSC ≒ 0.02 であれば十分な適合性を示しているといえる．SLSC ＞ 0.03 であれば，他の分布形へのあてはめを試みた方がよい」
「中小河川計画の手引き (案)」(1998)，気象庁異常気象リスクマップ: SLSC ≦ 0.04 を採用 (基準値を 0.02 や 0.03 にすると適合と判定される分布関数が見つからないケースがあったため，許容範囲を 0.04 まで広げたとのこと).
補足: SLSC はデータ年数 N に依存する (N が大きいほど減少) (葛葉，2010; 藤部，2011)
・適合度評価はどこまで必要か?

6-5 地域頻度解析 (Regional frequency analysis)
複数の地点のデータを合わせることにより，極値統計の推定精度を高める試み.

■ Station-year method
二重指数確率紙による graphical な方法を，複数地点に拡張したもの.
各地点のデータを，下記のように規格化してから確率紙に描きこむ.

$$y_i = \frac{x_i - M_i^{(2)}}{M_i^{(10)} - M_i^{(2)}} \tag{6-17}$$

x_i は地点 i の年最大日降水量，$M_i^{(2)}$ と $M_i^{(10)}$ は，地点 i の 2 年・10 年再現降水量 (あらかじめ地点ごとに Gumbel 分布を適用して求めておく).

■ Hosking の方法
クラスター分析で地点をグループ分けした後，グループごとに各分布関数の適合度を評価し，最適の分布関数を適用する.

■ 石原の方法 (都道府県単位)
Hosking の方法を簡略したもの．都府県を 1 グループとする.

■ その他，考えられる方法
高次の moment (2 次よりも 3 次) ほどデータの確率変動に影響されやすい．これに対応するパラメーター (GEV の場合，κ) の不確実性が，再現期間の推定誤差の大きな要因になる.
そこで，このようなパラメーターに限って広域平均値を使う (一方，α や β は地点ごとに求める) という方法も考えられる.

6-6 異常値の問題 (Outliers)
■ 異常値の例
・1896 年 9 月，彦根：日降水量 596.9mm
彦根 1 地点のデータに Gumbel distribution を当てはめると，再現期間 10 億年，鈴木・菊地原 (1984) の Station-year method でも 200 万年余.

九州の南に台風があり，本州を前線が縦断･･･総観的には集中豪雨の典型的パタン

・1950 年 8 月，苫小牧：日降水量 447.9mm
南東風場の局地豪雨.
類似の豪雨は，胆振沿岸で過去に数回発生.

■ 異常値をどう見るか

・極値統計手法の限界を反映
統計手法：定常・一様場を前提　⟷ 現実は変動・不均一場

・彦根豪雨の真の再現期間はいくらか？　気候モデルで 1 万年ランができれば･･･

■ 降水量の物理的上限はいくらか?
降水量の世界記録・日本記録
時間スケールとの関係

■ その他の話題

・大雨の極値パラメーターの広域分布
時間スケールとの関係
時間スケールが長い降水ほど，L-CV や L-skewness が増加 (κが減少)
　＝年々変動や sporadicity 増加
短時間降水は，緯度が高いほど L-CV や L-skewness が増加 (κが減少)
　＝年々変動や sporadicity 増加

・強風統計と地形効果

6-7 閾値解析，POT 解析 (Peaks-over-threshold analysis)
年最大値ではなく，日々の全観測値の中から上位値を取り込む.

■ 概要

$x \rightarrow$大で，GEV は一般化パレート分布 (generalized Pareto distribution; GPD) に漸近する.

$t \rightarrow 0$ で，$e^{-t} \rightarrow 1-t$ であることから,

$$F(x) = \exp[-\{1 - \frac{\kappa(x-\beta)}{\alpha}\}^{1/\kappa}] \rightarrow 1 - [1 - \frac{\kappa(x-\beta)}{\alpha}]^{1/\kappa}, \quad (\kappa \neq 0) \tag{6-18}$$

$$F(x) = \exp\{-\exp(-\frac{x-\beta}{\alpha})\}, \rightarrow 1 - \exp(-\frac{x-\beta}{\alpha}), \ (\kappa = 0) \tag{6-19}$$

指数分布 (exponential distribution).

閾値 u $(>\beta)$ のデータだけを考える (すなわち，$F(u)=0$ となるように CDF を書き直す) :

$$F(x) = 1 - [1 - \frac{\kappa(x-\beta)}{\alpha}]^{1/\kappa}] \rightarrow 1 - [1 - \frac{\kappa(x-u)}{\alpha^*}]^{1/\kappa}, \quad (\kappa \neq 0) \tag{6-20}$$

$$F(x) = 1 - \exp(-\frac{x-\beta}{\alpha}) \rightarrow 1 - \exp(-\frac{x-u}{\alpha^*}), \quad (\kappa = 0) \tag{6-21}$$

ただし $\alpha^* = \alpha - \kappa(u-\beta)$

■ 再現期間 T と F の関係

$$T(x) = 1/M(1-F(x)) \tag{6-22}$$

ただし，M は u を超える観測値の年間回数.
M を与えたとき，α^* と u は

$$\alpha^* = \alpha M^{\kappa} \tag{6-23}$$

$$u = \beta - \frac{\alpha}{\kappa}(M^{\kappa}-1) \tag{6-24}$$

■ POT 解析の実用的側面

1 年に複数回の大雨があったとき，そのすべてを使える (⟷年極値解析は最大値のみ).

ただし，「大雨の当たり年」のような気候変動要因に影響されやすい可能性はある? (年極値解析を同時に行い，結果を比べてみるとよい)

閾値の取り方の問題

複数の分布関数の中から適合するものを選ぶという手法は，一般的ではなさそう.

Ⅶ 気象観測とデータ

7.1 気象庁の気象観測 (Observation of JMA)

7.1.1 気象庁の組織 (organization)

■ 本庁 Main office in Tokyo

総務部 Adminstration Department

予報部 Forecast Department

短期予報，週間予報，台風情報など

観測部 Observations Department

観測，統計

地球環境・海洋部 Global Environment and Marine Department

季節予報，気候情報・気候変動，海洋，大気汚染情報

地震火山部 Seismological and Volcanological Department

■ 地方機関

管区気象台・沖縄気象台 Regional headquarters (6)

地方気象台 Local Meteorological Offices (50)

測候所 Weather Stations (2)

その他の付属機関

気象研究所 Meteorological Research Institute

気象大学校 Meteorological College

気象衛星センター Meteorological Satellite Center

・1990 年代後半から測候所の無人化

→ 特別地域気象観測所　95 ヶ所 Special automated weather stations

7.1.2 気象庁の観測網 (observation network)

■ 地上観測網とその変遷

・気象台・測候所・特別地域気象観測所

函館 1872 年，東京 1875 年，19 世紀末までに 80 地点超

気圧，気温，湿度，風，降水，日射・日照，視程など

・アメダス（地域気象観測所）Automated Meteorological Data Acquisition System

降水量 1300 地点

気温・風・日照時間 850 地点

積雪深 300 地点

・航空官署 Aviation weather service centers, Aviation weather stations

・区内観測 Local observation

アメダス以前の有人（委託）観測

1 日 1 回，日最高・最低気温と日降水量

7.2 気象庁の統計業務

7.2.1 平年値 (climate normals) について

■ 平年値の定義：30 年間の平均値

A 30 year period is used, as it is long enough to filter out any interannual variation or anomalies, but also short enough to be able to show longer climatic trends.

(https://www.wmo.int/pages/themes/climate/climate_data_and_products.php)

■ 歴史：
・1935: 国際気象機関 (International Meteorological Organization) の会議で 1901 ～ 1930 年の 30 年間を平年値の統計期間とすることが勧告された。
・1956: 世界気象機関 (World Meteorological Organization; WMO) が平年値を 10 年ごとに更新するよう勧告した。
・日本では：
1925 年の「理科年表」に 1886 ～ 1920 年の 35 年平均値を掲載.
1931 年，中央気象台が気候表「The Climate of Japan」を刊行。1897 ～ 1926 年の 30 年平均値を掲載。

■ 要注意点：「平年」という年があるわけではない。あるのは「平年値」。

■ 日別平年値の求め方
日々の不規則変動を除くため，9 日移動平均を 3 回行う。
$\sigma = \sqrt{20}$ の Gauss 関数 (正規分布関数) による平滑化とほぼ同じ。

■ 地球温暖化と平年値の問題
気候変動のもとで，平年値は「現在の気候の標準状態」を必ずしも表さない。
"normals may not represent the current state under a changing climate"
"no longer valid under a changing climate"

■ 「階級」と平年値についての注意

7.2.2 気象統計における官署移転等の扱い (site change)
■ 観測所が移転したとき，その前後の観測値をどう扱うべきか？
・1960 年ごろまでは原則として何もしなかった。
・1960 年ごろ，気象庁内で検討（斎藤錬一氏）．→統計接続，統計切断。
・移転する官署が増えるにつれ，統計切断が増えてきた。極値記録の抹消。

■ 現在の方法：
・気温等については，移転前のデータを補正して統計する。極値は切断しない。
・気温の補正方法
主成分分析による広域成分の抽出，階段関数による重回帰
階級別日数（真夏日，熱帯夜など）の補正
補正式を過去に遡って適用し，階級別日数を数え直す

■ 東京の観測所の履歴
・大手町から北の丸公園への移転 2014 年 12 月 2 日
・移転の背景
気象庁本庁の移転　2020 年，虎ノ門庁舎落成予定。
移転に伴う観測値の変化→ 3 年間の同時観測

7.2.3 気象庁における観測データの品質管理 (quality control)

■ 気象官署・アメダスの観測値の管理

・常時監視

基準を超える「異常」があったときは，直ちに状況を調査し，必要に応じてデータを無効化する。

・事後の品質チェックを行う。

■ 観測所の維持管理。定期的な草刈りなど。

・雑草対策：

いろいろな方法（草刈り，薬物，防草シート）

防草シートの問題点：蒸発の抑止→ 顕熱増加の可能性?

■ 誤データが出る理由：

測器の故障，設定ミス

自然要因（雪の付着・凍結，雷災など）

動植物（虫，クモ，鳥，落ち葉，雑草類）

人為的要因（野焼き，アイドリングなど）

多種多様→ 1 つ 1 つ対応するしかないのが実情。

7.2.4 観測データの提供と電子化 (digitization)

■ 電子データの提供

気象庁ホームページ

電子ファイル… 気象業務支援センターによる提供

■ Data rescue:

・背景：データの危機的状況，電子媒体の発達

・日本の取り組み

気象庁の取り組み：日・時別降水量

区内観測データの電子化

7.3 地上気温観測に関わる事項 (surface observations)

7.3.1 気温の観測方法 (temperature observation)

■ 温度計の設置

・地上高

日本では 1.5m，明治時代は 1.2m

国際的には 1.25m ～ 2m

・気温変動の平均化

温度計の応答時間　数十秒　気象官署と一般アメダスで違う

10 秒値を 6 回平均したもの＝正 10 秒値

正 10 秒値のうち終点が 00 秒のもの＝正 1 分値

・気象官署とアメダスの違い

気象官署もアメダスを兼ねている。現在は観測値は同じ。

以前は観測値が違った different records from the same thermometer (synoptic observation and AMeDAS)
← 温度計は同一。データサンプリングの時間差 difference in the timing of data sampling

■ 温度計の変遷 (history of thermometers)
・1960 年代まで百葉箱 instrument screen ＋棒状温度計 bar thermometer , manual observation
・1970 年代からは通風筒 ventilated sheild ＋抵抗温度計 resistance thermometer, automated observation

■ 積雪時の気温観測 (observation on snow)
・積雪があるときは，雪面からの高さ 1.5m の気温を測ることになっている。
しかし，温度計の高さを積雪に合わせて調節するのは実際には無理
→昔は雪かきをした
今は冬の前に温度計をかさ上げしておく（雨量計も）。

7.3.2 観測環境の問題 (observational environments)
■ 都市バイアスの問題 (urban bias)

■ 日だまり効果 (reduced ventilation)
防風ネット周囲の気温分布 Temperature distribution near a windbreak net
気象庁構内観測 An observation study in the JMA site
・遮蔽物 (樹木) があるところの気温
・アスファルト道路の影響 (effect of road surface)
・江川崎「41.0 ℃」の信頼性 (reliability of the Japan's TPTP record of 41.0 ℃)

7.3.3 観測時刻の問題 (observation time)
■ 観測の時間間隔 sampling intervals
・1 時間値，10 分値，連続値
・観測時刻
戦前は 4 時間単位（06, 14, 22 時＋ 02, 10, 18 時）
1953 年以降は 3 時間単位（03, 09, 15, 21 時＋ 06, 12, 18, 24 時）
・日平均・最高・最低気温の定義の問題
観測回数，日界の問題
日界の違い observation time of the day

7.4 他の気象要素の観測に関する問題
7.4.1 降水量観測 (precipitation observation)
■ 降水量とは何か

■ 雨量計に関する話題 (raingauges)
・雨量計の変遷

貯水型から転倒ますへ
観測単位変更に伴う降水日数の補正
・捕捉率の問題
・転倒マス雨量計の特性：非常に強い雨への追随

■ 降水量観測に関するその他の話題
・感雨器導入とその影響
・雪による障害
・降水量観測に対する遮蔽物（樹木）の影響

7.4.2 風の観測 (wind observation)
■ 風観測の基本
地上風の性質，変動特性
平均風速と瞬間風速，突風率
風速の高さ分布と観測高度の問題
風速計の設置状況

■ 風速計の変遷 (history of anemometers)
4 杯型→ 3 杯型→風車型

■ 風観測に関するその他の話題
雪による凍結障害
風観測に対する遮蔽物（樹木）の影響

7.4.3 その他の要素
・日照計の変遷
・雲量観測に対する月光の影響，その経年変化
・積雪と降雪観測
・ゾンデ観測：日射補正，紐の長さ，ゾンデの種類など