まえがき

この資料では，主として大気海洋に共通する統計データ解析手法を解説する．これらの手法の多くは他分野でも用いられているが，いくつかの手法(特異値分解解析，特異スペクトル解析)はもっぱら大気海洋分野で利用されている．もちろん，大気・海洋それぞれにのみ適用できる，より物理的なプロセスと密接に関係した解析方法でも多いが，通常それらの手法は統計データ解析の範疇には入らず，この資料でも対象としていない．

これらの手法を整理する仕方としてこの資料では，解析の対象とするデータの形態から，**単変量(univariate)**解析，**２変量(bivariate)**解析，**多変量(multi-variate)**解析として分類した．ただし，異なる分類に入っている手法でも相互につよく関係する場合があり，注意が必要である．例えば，単変量解析のスペクトルと２変量解析のコヒーレンスは非常に近い関係がある．また，特異スペクトル解析は，その目的は名前に含まれているスペクトルと同様に周期成分の検出であるが，手法の観点からは延長した経験的直交関数展開とほぼ同じである．

百聞に一見はしかずという通り，目で見ることは理解を促進する．おそらく理解とは，純粋に知的な理解と感覚的な体得とが一体になったものであろう．後者はやってみなくては身につけることはできない．この資料では，実際に Matlab（もしくは Octave）を使って学習者自体が計算を行い，図を表示することによって，理解の深化と，体得への一歩を踏み出すことを意図している．また，Matlab を用いた演習を行なうことで，学習した内容を各人の興味にしたがって研究にも生かすことがより容易であろう．

なお，以下では青は定義を，赤は重要なポイントを，緑は Matlab/Octave での関数名を示す．

1. 確率・統計の基礎

この章では，大気・海洋の統計データー解析で用いられる手法や概念の基礎となる，一般的な確率・統計の事項を整理する．これらの事項の証明は，確率・統計の書籍に記載されているので，多くの場合この資料中で証明を述べることはせずに，それらの書籍を示している．証明が必要である場合には，それらの書籍を参照されたい．

1.1. 母集団と標本

統計解析の対象のほとんどは，数値である．例えば，地球の平均気温や，太陽黒点の数は数値として表される．まれになんらかの定性的性質も統計解析の対象として扱われるが，この資料では扱わない．

統計解析では多くの場合，**母集団(population)**と**標本(sample)**とを明確に区別して考える必要がある．母集団とは対象とする数値の集団全体であって，標本はその一部である．母集団を特徴付ける数値は**パラメータ(parameter)**と呼ばれ，通常ギリシャ文字で表される．例えば，母集団の平均（母平均 population mean）は μ であり，分散(母分散 population variance)は σ^2 である．また母集団パラメータの推定値を，＾記号をつけて，$\hat{\mu}$ のように表す．標本から計算される数値は**統計量(statistics)**と呼ばれ，通常英字で表現する．例えば，標本平均(sample mean)は $\overline{x}$ であり，標本分散(sample variance)は s^2 で示される[1]．なお，英語では，**単数扱いの statistics** は統計学を，**複数扱いの statistics** は統計量を表すので注意しよう．母集団のパラメータは一つの値しか持たないが，標本から求められた統計量は標本のセットが違えば当然異なり，真の値である母集団のパラメータとも異なる．統計がからむ科学では，いかに限られた標本から信頼できる母集団のパラメータを推定できるかが，鍵となる問題である．

1.2. 確率

今問題とする変数が，なんらかの値をとるとしよう．この値は，さいころの様にとびとびの値をとる，すなわち**離散的(discrete)**でも良いし，温度のように**連続的(continuous)**でもよい．また，１から６までと上下限があってもよいし，絶対温度のように下限はあるが上限はなくてもよいし，上下限ともなくてもよい．このような変数が，ある値をことを試行(trial)とよび，取りうる結果を事象(event)といい，事象の全体を全事象または根源事象(elementary event)と呼ぶ．全事象をしばしば記号Ωで表す．trial，event という言い方は，さいころを投げるような一回２回と数えられる状況に相応しいが，地球の気温のように時間・空間的に連続的な分布についても利用してもよい．また問題としている変数は，**確率変数(stochastic variable)**と呼ばれる．stochastic は辞書では確率的なという意味となっているが，「確率的に取り扱わなくてはならない」という気分ではないかと思う．

---

[1] あるパラメータについて母集団の推定値と，標本から計算した値が，なにが違うのかという疑問が生ずるなら，いいセンスをしている．多くの場合には違いはなく，あえて母集団の推定値と標本からの計算値を区別する場合には後者が前者のなんらかの意味でよい推定値であない場合であろう．

全事象のうち, 特定の事象（さいころが 1 の目を出す）または, 特定の事象の範囲（温度が 10~11 度の範囲である）が，生ずる（頻度の）割合が**確率(probability)**である．特に連続的な値をとる確率変数については，特定の事象を取る確率はゼロであり，値の範囲を必ず考える必要がある[2].

確率の定義の元となる「生ずる割合」は，無限回の試行をしなくては知ることができない，つまり実際に知ることはできない理想的な値である．現実のさいころは重さに多少の偏りもあるだろうし，面と面が正確に直角でもないだろう．したがって，ある特定のさいころについての出目の確率を，正確に知ることはできないが，実用十分な範囲で 1/6 と仮定するのである．

1.3. 確率密度関数(pdf)と累積確率関数(cdf)

離散的な確率変数 X が実現値 x_i を取る確率を，**確率関数(probability function)**と呼び,

$$P(X = x(i)) = W_{x(i)} = W(i) \tag{1.1}$$

と書く．ここで P は確率を表し，W は確率関数である．確率関数の独立変数は，事象 x_i（さいころなら 1,2,...,6）である. 理想的なさいころの確率関数は, 単一の値 $P(1)=P(2)=...=P(6)=1/6$ を取る．

連続的な確率変数 X については，すでに述べたように，実現値がある値を取る確立は常にゼロで意味がなく，実現値がある区間に入る確率に基づいて議論する必要がある．そこで X の実現値が，$x \sim x+\Delta x$ の間に入る確率を区間幅 Δx で割ったものを，確率密度(probability density)または**確率密度関数(probability distribution function, PDF)**と呼ぶ．すなわち,

$$P(x < X \le x+\Delta x) = W(x)\Delta x$$

である．ここで，W が確率密度関数である．確率密度関数は必ず正の値をとる．

また離散・連続を問わず，X の実現値がある値 x 以下である確率

$$F(x) = P(X \le x) \tag{1.2}$$

を，分布関数(distribution function)または**累積分布関数(cumulative distribution function, CDF)**と呼ぶ．またまれに累積確率(cumulative probabilities)と呼ばれることもある．明らかに，PDF と CDF の間には，次の関係が成り立つ.

$$F(x) = \int_{-\infty}^{x} W(x)dx \tag{1.3}$$

$$W(x) = \frac{d}{dx}F(x) \tag{1.4}$$

また，PDF の区間 $(-\infty,\infty)$ での積分値は 1 となるので，$x \to \infty$ の極限で CDF は 1 である．

統計検定・推定では，しばしば CDF の逆関数が必要である．しばしば統計の本では，累積分布関数が α を与える，x などというもって回った言い方が登場することごあるが，数学的に普通に言えば，$x = W^{-1}(\alpha)$ という関係であるにすぎない．この CDF の逆関数には，PDF や CDF とい

[2]例えば気温は正確に 10 度になることはなくその確率はゼロであるけれど，10 度～11 度の範囲の値を取る確率は一定の大きさを持つ．なお，有限桁で表現される場合には，その桁数の範囲で正確に 10 度になることはありえて, その際には有限桁となったために離散化されていると考えることもできる．しかし，もともとが連続的である現象については，連続的な扱いをすることが一般的である．

う広く流通するコンパクトな名前は欧米でも付けられていないようであるが，重要な概念に用語を割り当てることは明晰な議論には重要なので，この資料では，**累積分布逆関数(inverse of cumulative distribution function, ICDF)**と呼ぶことにする．

確率関数あるいは確率密度関数が定まっていることを，確率分布が与えられているという．例えば，ある過程の確率が正規分布するとは，正規分布する確率関数を持つということである．なおしたがって，確率分布という用語は，確率関数・確率密度関数を規定するのに対して，分布関数という用語は（確率）分布という用語と似ているが，実際には確率関数・確率密度関数の積分であり，両者は混同しやすいので注意しよう．混同を避けるには，分布関数ではなく**累積**分布関数と記す方良い．この資料ではこれ以降，確率密度関数を PDF，累積分布関数を CDF と書く．

1.4. 平均と期待値

平均という用語は，母集団の平均にも，標本の平均にも使われるので，場合によっては混乱を招く．一方，**期待値(expected value)**は，離散的な確率変数に対して

$$E(X)=\sum_i P(x(i))x(i) \tag{1.5}$$

連続的な確率変数に対しては

$$E(X)=\int_\Omega P(x)x\,dx \tag{1.6}$$

と定義される．ここで，Ω は全事象を示している．期待値は X の平均に等しい．なお，**期待値が等しいのは母集団平均であって，標本平均ではない**ということは心に留めておこう．

1.5. 回帰係数と相関係数

二つの観測されたデーター列 $x(i), y(i), i=1,...,N$ の間の相互関係を，

$$y'(i)=ax'(i)+b \tag{1.7}$$

という傾き a で切片 b の線形関係で説明しよう．ここで x', y' は $x'(i)\equiv x(i)-\overline{x}, y'(i)\equiv y(i)-\overline{y}$ と定義される平均からのずれで，計算を簡単にするために導入している．説明される分散を最大（説明されない分散を最小）にする a, b は，標準的な最小二乗法によって次のように求めることができる．まず，観測値の推定値からのずれである，残差(residual)の二乗和は

$$\begin{aligned}\varepsilon &= \sum_{i=1}^{N}(ax_i+b-y_i)^2 \\ &= \sum_{i=1}^{N}\left\{a^2x_i'^2+b^2+y_i'^2+2abx_i'-2ax_i'y_i'-2by_i'\right\}=\sum_{i=1}^{N}\left\{a^2x_i'^2+b^2+y_i'^2-2ax_i'y_i'\right\}\end{aligned}$$

である．最後の等号では偏差の和がゼロ($\sum x_i'=\sum y_i'=0$)という関係を用いた．この εの，a についての偏微分と，b についての偏微分はそれぞれ

$$\frac{\partial}{\partial a}\varepsilon=2\sum_{i=1}^{N}\left(ax_i^2-x_iy_i\right),\quad \frac{\partial}{\partial b}\varepsilon=2\sum_{i=1}^{N}b$$

であって，説明される分散が最大になるためには，これらの微分はともにゼロであるから，明らかに b=0 であり，また

$$a = \frac{\sum_{i=1}^{N} x'(i)y'(i)}{\sum_{i=1}^{N} x'(i)^2} = \frac{\sum_{i=1}^{N} (x(i)-\overline{x})(y(i)-\overline{y})}{\sum_{i=1}^{N} (x(i)-\overline{x})^2} \tag{1.8}$$

となる．この a を**回帰係数(regression coefficient)**と呼ぶ．したがって，最大の分散を説明する線形関係を，平均のずれではなく x, y 自体について表せば，

$$y_e(i) - \overline{y} = a\left(x(i) - \overline{x}\right) \tag{1.9}$$

である．ここで，下付き添え字 e は，推定値(estimated value)であることを示している．

回帰係数は(物理量 yの単位/物理量 x の単位)という単位を持つことが(1.8)から, 明らかである．したがって，異なる物理量間の回帰係数，例えばエルニーニョとある地点の気温・降水量の関係がどちらが強いかを回帰係数で比較することは意味がない．このような目的のためには，回帰係数を無次元化・規格化すればよいということが，思い浮かぶだろう．(1.8)を無次元化するには視察により，

$$r = \frac{\sum_{i=1}^{N} (x(i)-\overline{x})(y(i)-\overline{y})}{\sqrt{\sum_{i=1}^{N} (x(i)-\overline{x})^2}\sqrt{\sum_{i=1}^{N} (y(i)-\overline{y})^2}} \tag{1.10}$$

とすればよいことが見て取れる．この係数が，**相関係数(correlation coefficient)**である．相関係数は２変量・多変量解析の基礎となる係数で，しばしば r または ρ で表される．また，x と y それぞれの不偏分散を S_x^2, S_y^2 と書けば，相関係数は，

$$r = \frac{1}{N-1}\sum_{i=1}^{N} \frac{(x(i)-\overline{x})(y(i)-\overline{y})}{S_x S_y} = \frac{1}{(N-1)S_x S_y}\left(\sum_{i=1}^{N} x(i)y(i) - N\overline{xy}\right) \tag{1.11}$$

とも表される．相関係数を用いることで，回帰係数を

$$a = r\frac{S_y}{S_x} \tag{1.12}$$

と表すことができる．また相関係数は，共分散(covariance)

$$C_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}\left(x(i)-\overline{x}\right)\left(y(i)-\overline{y}\right) \tag{1.13}$$

を標準偏差で規格化(normalize)したものでもある．

相関係数は±1 の範囲の値を取り，値の絶対値が大きいほど２変数間に強い線形関係があることを示している．なお，２変数間の関係が線形ではなく，例えば２次曲線の場合には，非常につよい関係があっても相関係数は低いということがあり得る. このため, 相関係数の推定に先立って，散布図をチェックすることが望ましい.

回帰係数を用いると推定値が得られるだけではなく，観測値の推定値からの**残差(residual)**が得られることも重要な場合である．そこで，ここで残差の性質を説明しよう．まず，観測値の推定

値からの残差, $y_r(i)$,を

$$y_r(i) = y(i) - y_e(i)$$

と定義する．推定値の平均は(1.9)から，明らかに観測値の平均に一致し，従って残差の平均はゼロである．

また，推定値と残差は無相関となることが次のように示される．

$$\begin{aligned}
\frac{1}{N-1}\sum_{i=1}^{N}\left[y_r(i)\right]\ \left[y_e(i)-\bar{y}\right] &= \frac{1}{N-1}\sum_{i=1}^{N}\left[y(i)-r\frac{S_y}{S_x}\left(x(i)-\bar{x}\right)-\bar{y}\right]\left[\left\{r\frac{S_y}{S_x}\left(x(i)-\bar{x}\right)+\bar{y}\right\}-\bar{y}\right] \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left[\left(y(i)-\bar{y}\right)-r\frac{S_y}{S_x}\left(x(i)-\bar{x}\right)\right]r\frac{S_y}{S_x}\left(x(i)-\bar{x}\right) \\
&= r^2S_y^2 - r^2S_y^2 = 0
\end{aligned}$$

したがって，観測された分散は，次のように，推定値の分散と残差の分散の和で表される．

$$\begin{aligned}
S_y^2 &= \frac{1}{N-1}\sum_{i=1}^{N}\left[y(i)-\bar{y}\right]^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left[y_e(i)+y_r(i)-\bar{y}\right]^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left[\left(y_e(i)-\bar{y}\right)+\left(y_r(i)\right)\right]^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left(y_e(i)-\bar{y}\right)^2+\frac{1}{N-1}\sum_{i=1}^{N}y_r(i)^2
\end{aligned}$$

なお一般に，データーを直交する成分の和で表現する場合に，分散は各々の成分の分散の和で表される．上式の推定値の分散は，

$$\frac{1}{N-1}\sum_{i=1}^{N}\left(y_e(i)-\bar{y}\right)^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(r\frac{S_y}{S_x}\left(x(i)-\bar{x}\right)\right)^2 = r^2S_y^2$$

であるので，結局，観測された分散の内，**回帰関係で説明できる分散の割合が r^2 で，説明できない分散の割合が 1- r^2 である**．この著しい性質があるために，相関係数が 0.5, 0.6, 0.7 であれば説明できる分散の割合は各々1/4, 1/3, 1/2 となる．従って，相関係数が 0.7 を上回る場合は，支配的(dominant) という表現を使うことができる．

なお気候変動を議論する上で，しばしば，上の回帰関係から得られる残差を利用して，注目している現象から，他の減少の効果を除く，という操作が行われる．例えば，アリューシャン低気圧の強さと札幌の気温とに有意な相関関係が検出されたとしよう．この関係が，経年時間スケールで最もエネルギーの強いエルニーニョに関連して生じているのか，あるいは独立に生じているのかは興味深い問題である．この問題を扱うために，アリューシャン低気圧の強さと札幌の気温の両方から，エルニーニョによって生ずる成分を各々の回帰関係を用いて除いてみよう．そして，エルニーニョの効果を除いた残差同士で再び相関を計算して，もし有意であるなら，「アリューシ

ャン低気圧の強さと札幌の気温の間の有意な相関は，エルニーニョに付随して生じているのではない.」と結論できる.

1.6. モンテカルロ法の概説

モンテカルロは，フランス南東部・モナコ公国北東部の観光保養地で，国営賭博場で有名な街であり，モンテカルロ法という名称は，ルーレットのようなランダムな方法を暗に示している．**モンテカルロ法(Monte-Carlo method)**とは，コンピューターによって生成されるランダム変数を用いて，何らかの解を得る方法である．例えば，モンテカルロ法によって，ランダム確率変数から求められる統計推定量の分布を用いて，ある仮説を棄却できる有意水準などを推定することができる．また，この資料とはかかわりがないが，乱数を用いて面積を計算する方法もモンテカルロ法である．モンテカルロ法ではなく，モンテカルロ・シミュレーション(Monte-Carlo simulation)とも言うが，両者の間にそれほど厳密な区別はない．多分シミュレートには，なにかを“なぞる”という気持ちが入るので，その気持ちの強弱によって使い分けると良いだろう．

様々な統計的な推定および検定をモンテカルロ・シミュレーションによって行うことができる．理論的に推定・検定を行える場合も多いが（というよりも理論的な推定・検定が使えそうなら，それを使うように話を持っていくのが普通である），理論的な推定・検定法が与えられていない場合でも，一般にモンテカルロ・シミュレーションを使えば推定・検定ができる．また，理論的な推定・検定法が与えられている場合でも，どこまで理論が妥当であるのかを調べるためにも，モンテカルロ・シミュレーションを使うことは非常にメリットがある．例えば，ごく普通に使われている相関係数やコヒーレンシーに関する有意性検定の導出は，実は非常に難しくほとんどの教科書では示していない．ほとんどの研究者はその導出を追う必要はないが，もし妥当性を確認する必要があれば（例えば論文の査読者が要求すれば）モンテカルロ法で確認することができる．

モンテカルロ法を用いるためには，必ず乱数生成が必要である．Matlab では，**randn** が正規分布する乱数を，**rand** が(0,1)の区間の一様乱数を生成する．

演習問題 1.1 区間(-1, 1)の間の n 個の一様乱数を m セット生成し，セットごとに平均して，その平均値の確率密度関数(PDF)をモンテカルロ法によって求めよ．n=100 に固定し，m=100, 1000, 10000 の 3 通りについて図示せよ．なお，頻度分布を求めるには，**hist** 関数を用いると良い．ベクトル v に頻度分布を求めるべきデータが格納されていて，例えば x=-0.3:0.02:0.3;(x=-0.3, -0.28, -0.26…)を中心を持つ bin における頻度分布を vh というベクトルに格納するには, vh=hist(v,x)とすればよい.

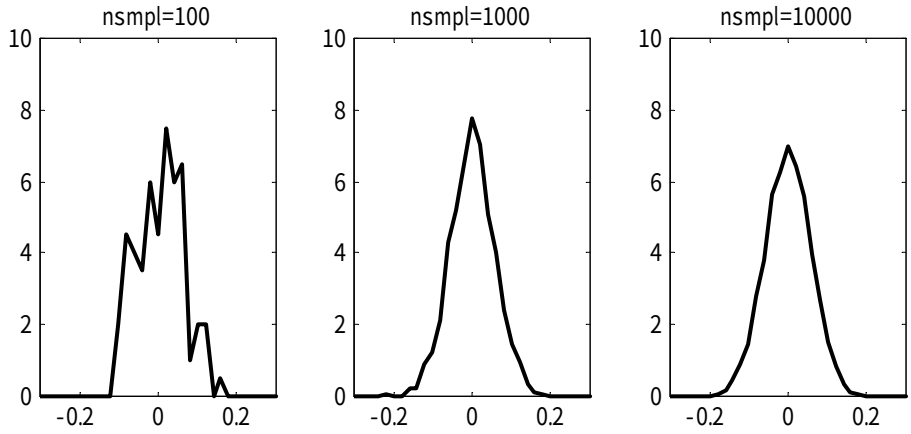

図 1.1　演習問題 1.1 の回答．一様乱数の一定の数の平均値の分布は，セットの数を増やせば，後に示す大数の法則によって正規分布に漸近する．

1.7. よく使われる確率分布とその Matlab 関数

統計に出てくる分布には，大別すると２通りある．一つは，現象の振る舞い自体についての分布で，２項分布・ポアソン分布・正規分布がその代表格である．他の一つは，現象自体ではなく，それに何らかの数学的操作を加えた統計量の分布であって，χ^2(カイ二乗)分布・t 分布・F 分布がその代表格である．なお, t は小文字で, F は大文字で書くのが慣例となっている．χ二乗分布は，２乗平均および分散の推定，さらにスペクトルの有意性判定に用いられる．t 分布は，平均値・平均値の差の推定，さらに相関係数の有意性の検定に用いられる．F 分布は分散比の推定・検定に用いられる．

Matlab では Statistical Toolbox で，Octave では statistics/distribution で各種の確率密度分布，累積確率密度分布，累積確率密度分布の逆関数が提供されている．ただし自由度は整数に限定されていることが多く，高度な解析ではユーザー独自に自由度を実数に拡張する必要がある．

表 1.1 統計推定･検定に用いられる主要な確率分布に関する関数. 上段は Matlab の statistics Toolbox で，下段は Octave の statistics/distribution で提供されている関数名である．引数を PDF の列のみ示しているが，CDF は同じであり，ICDF の場合は x の代わりに CDF が入る．normpdf の[,mu,sigma] は省略することができ，省略するとそれぞれ 0 と 1 が仮定され標準正規分布となる．

	確率密度分布 (PDF)	累積分布関数 (CDF)	累積分布逆関数 (ICDF)
正規分布	normpdf(x[,mu,sigma])) normal_pdf	normcdf normal_cdf	norminv normal_inv
χ二乗分布	chi2pdf chi2_pdf	chi2cdf chi2_cdf	chi2inv chi2_inv
F 分布	fpdf(x,N1,N2) f_pdf	fcdf f_cdf	finv f_inv

t 分布	tpdf(x,N) t_pdf	tcdf t_cdf	tinv t_inv

1.7.1. 正規分布

正規分布(normal distribution)は統計解析において最も広く使われる分布で，**ガウス分布(Gaussian distribution)**とも呼ばれる．正規分布の確率密度関数(pdf)はガウス関数となり，

$$W(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\} \tag{1.14}$$

で定義される．正規分布の平均はμ，標準偏差はσである．逆に言えば，ガウス関数を標準偏差がσになるように幅をきめ，中心をμとし，$(-\infty,\infty)$での積分が 1 になるように規格化したものが正規分布である．この正規分布を，$N(\mu, \sigma^2)$と表示することも多い．平均が 0，標準偏差が 1 である正規分布 N(0, 1)を，**標準正規分布(standard normal distribution)**という．

正規分布が有用であるのは，正規分布ではない確率過程の平均値も，往々にして正規分布となるためである．この性質は数学的には**中心極限定理**として知られる，「同一の確率分布に従う互い独立な N 個の確率変数 $x(1), x(2), \ldots, x(N)$の和（平均でも定性的には同じ）は，$N\to\infty$の極限で正規分布となる」[3]．例えば，１時間あたりの降水量分布はおそらく正規分布してはいないだろう．しかし，月平均すると正規分布に近くなることが期待される．このように，多くの大気・海洋の変動は正規分布がまず考えるべき確率分布である．ただし一方で正規分布しない分布が多いことにも注意しなくてはならない．正規分布ではない分布としては，ゆがんだ分布が生じる場合や，確率密度が２つの峰を持つ bi-modal と呼ばれる分布などがある．例えば，乾燥地域の降水量分布は，正規分布ではない場合が多い．正規分布は多くの統計解析の理論で仮定されているので，正規分布ではない分布については，標準的な統計解析理論の適用に注意を払わなくてはならない．

中心極限定理に関連して，より簡単な例を紹介しておこう．N 個のデータ $x(1), x(2), \ldots x(N)$が正規分布 $N(\mu, \sigma^2)$に独立に従うとき，標本平均 $\overline{x} = \sum_{i=1}^{n} x(i)/n$ は正規分布 $N(\mu, \sigma^2/N)$に従う．つまり母集団分布が正規分布であれば，標本平均は母集団分布よりも小さな分散を持つ正規分布に従い，分散は N だけ小さくなる．なお，正規分布を示す N と，標本の個数である N とは紛らわしいが文脈で区別して欲しい．

Q. ある量の誤差が正規分布することが分かっている（平均はゼロ）。誤差の大きさを正規分布の標準偏差 σで表す。n 回の測定を行ってその平均を求める場合に、得られた平均値の誤差の標準偏差を答えよ。

A. 上の理論から、平均値の誤差は $N(\mu, \sigma^2/n)$に従うので、その標準偏差は $\sqrt{\sigma^2/n}$ である。

[3] 証明は柴田 1996

対数正規分布(log normal distribution)は，確率変数の値自体(x)ではなく，その対数($\log(x)$)が正規分布に従う．例えば，誤差の蓄積を考える場合に，個々の誤差が正規分布に従い，その誤差が足し合わさって最終的なの誤差となるのであれば，最終的な誤差は正規分布に従うであろう．一方誤差が掛け合わされるのであれば，最終的な誤差は対数正規分布に従うことが期待される．

1.7.2. χ^2(カイ二乗)分布

標準正規確率分布 N(0, 1)に従う N 個の確率変数 $x(1), x(2), \ldots x(N)$に対して，$Y = \sum_{i=1}^{N} x(i)^2$ が従う分布を，自由度 N の **χ^2（カイ二乗）分布**という[4]．自由度 N のカイ二乗分布の確率密度関数(PDF)は

$$W_{\chi^2}(y, N) = \frac{1}{2^{N/2}\Gamma(N/2)} y^{(N/2)-1} e^{-y/2} \tag{1.15}$$

で与えられる．ここで Γはガンマ関数(Gamma function)である．

ガンマ関数は

$$\Gamma(\lambda) = \int_0^\infty x^{\lambda-1} e^{-x} dx \tag{1.16}$$

で定義される．ガンマ関数の性質で特に有名なのは，次の 2 式である．

$$\Gamma(\lambda+1) = \lambda\Gamma(\lambda) \tag{1.17}$$

$$\Gamma(1) = 1 \tag{1.18}$$

これら 2 式から，とくにλが整数であれば，$\Gamma(n+1) = n!$となる．また，1/2 のガンマ関数もしばしば使われ

$$\Gamma(1/2) = \sqrt{\pi} \tag{1.19}$$

である．

標準正規分布ではなく，一般の正規分布 $N(\mu, \sigma^2)$ に従う確率変数 x に対して，$Y = \sum_{i=1}^{N} (x(i)-\mu)^2/\sigma^2$ という変数をつくると，明らかに Y も自由度 N の χ^2 分布に従う．なお，母平均が未知であれば，標本平均で置き換えなくてはならないが，この場合は，$Y = \sum_{i=1}^{N} (x(i)-\bar{x})^2/\sigma^2$ が自由度 N-1 のカイ二乗分布に従う．自由度が 1 少ないのは，平均を求めるのに自由度 1 を使ってしまったからである(柴田 1996, p. 112).

また**自由度 N のカイ二乗分布の平均は N，分散は $2N$ である**[5]．この性質は，スペクトルの有意性検定で利用される．

---

[4]証明は柴田 1996, p. 69-78，もしくは小針, 1973 p. 71-72

[5]柴田 1996, p. 78-79

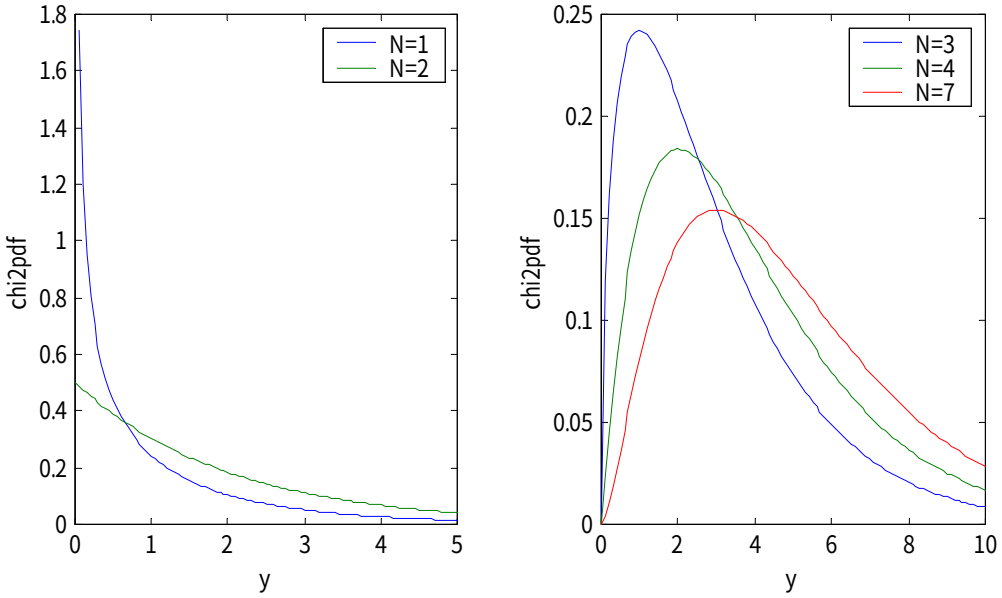

図 1.2　χ^2 分布の確率密度関数．原点では，N=1 で発散，N=2 では 0.5, N≥3 ではゼロである．また，N=1 と 2 は単調減少，N≥3 では上に凸である．

1.7.3. 非整数自由度のカイ二乗分布の累積分布逆関数

カイ二乗分布に従う確率変数について統計推定・検定を行うには，累積分布逆関数(ICDF)を求めることが必要である．Matlab では chi2inv で提供されている．ただし，chi2inv は整数自由度について実行可能となっていて，後に述べるようにスペクトル推定などで実数の自由度が必要となる場合がある．この場合，chi2inv では不十分である．そこで，カイ二乗分布の ICDF の計算方法を紹介しよう．

その道具として，ガンマ(Γ)密度関数を導入する．

$$w_\Gamma(x\,|\,a,b)=\frac{1}{b^a\Gamma(a)}x^{a-1}e^{-\frac{x}{b}}$$

この密度関数を，(1.15)のカイ二乗分布の PDF と比較すると，$a=N/2, b=2$ の場合にカイ二乗分布に一致する．つまりガンマ密度関数はカイ二乗分布を含んでいる．ガンマ密度関数の ICDF は Matlab では gaminv(p, a, b)で（多分）正の実数の a,b に対して提供されている．したがって，chi2inv(p,v)に代えて chi2inv_nint(p,v)を次のように定義すればよい．もちろん，整数自由度については chi2inv も chi2inv_nint も同じ結果を返す．

```
function icdf = chi2inv_nint(p,v);
% 非整数の自由度 v についても答えを返す．chi2 分布の累積密度関数の逆関数.
% 入力
%     p: 累積確率分布(0<=p<=1)
%     v: 自由度 (0<v)
% 出力
%     icdf: inverse of cumulative distribution function
```

```
% ちなみに statistical toolbox の chi2inv は，非整数の自由度 v については
% NaN を返す.
% 2001/12/10 見延 作
if isempty(p); error('*chi2inv_nint* p is empty'); end
if isempty(v); error('*chi2inv_nint* v is empty'); end
if sum(p<0); error('*chi2inv_nint* p が負です'); end
if sum(p>1); error('*chi2inv_nint* p が１以上です'); end
if sum(v<0); error('*chi2inv_nint* 自由度 v が負です'); end
if sum(v==0); error('*chi2inv_nint* 自由度 v がゼロです'); end
if (ndims(p)~=ndims(v)); error('p と v の次元が違います'); end
if (sum(size(p)~=size(v)); error('p と v の大きさが違います（次元は同じです）'); end

%ガンマ累積分布関数の逆関数を呼ぶ.
icdf = gaminv(p,v/2,2);
```

1.7.4. F 分布

カイ二乗分布に従う独立な二つの確率変数 Y_1, Y_2 を考え，それぞれの確率変数の自由度を N_1, N_2 とする．確率変数 $Y = (Y_1 / N_1)/(Y_2 / N_2)$ の分布は，自由度(N_1, N_2)の **F 分布(F-distribution)**と呼ばれる．

F 分布はまた，スネデカーの F 分布 (Snedecor's F-distribution)，またはフィッシャー分布(Fisher distribution) とも呼ばれる．自由度(N_1, N_2)の F 分布の PDF は，次式で与えられる[6].

$$W(y,N_1,N_2) = \frac{N_1^{(N_1/2)} N_2^{(N_2/2)} \Gamma\left(\frac{N_1+N_2}{2}\right)}{\Gamma(N_1/2)\Gamma(N_2/2)} \frac{y^{(N_1/2)-1}}{(N_1 y + N_2)^{(N_1+N_2)/2}} = \frac{N_1^{(N_1/2)} N_2^{(N_2/2)}}{B\left(\frac{N_1}{2},\frac{N_2}{2}\right)} \frac{y^{(N_1/2)-1}}{(N_1 y + N_2)^{(N_1+N_2)/2}} \tag{1.20}$$

ただし，y>0 についてで，y≤0 については W=0 である．ここで B はベータ関数で，次のように定義される．

$$B(a,b) = B(b,a) = \int_0^1 t^{a-1}(1-t)^{b-1}\,dt \quad .$$

なお，ベータ関数とガンマ関数は次の関係を持つ．

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{1.21}$$

[6]証明は，柴田 1996, p. 80-84，小針 1993, p.176-179

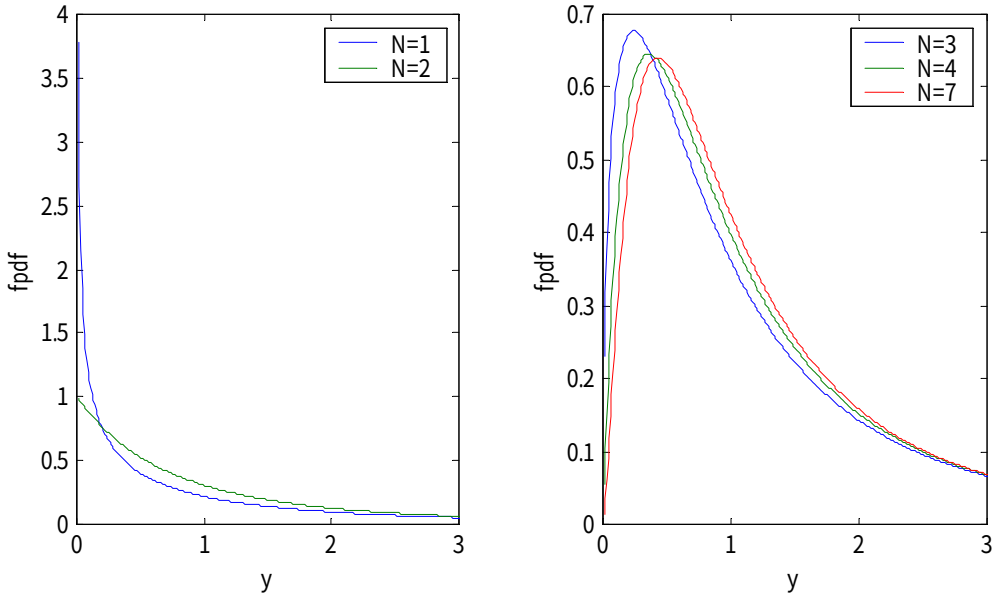

図 1.3. N_1=5 の F 分布の PDF．図中の N は N_2 を意味する．

1.7.5. *t*-分布

準正規分布 $N(0,1)$に従う確率変数 Z と，自由度 N のカイ二乗分布に従う確率変数 Y からなる確率変数 $T = Z / \sqrt{Y / N}$ は自由度 N の ***t* 分布(*t*-distribution)**に従う．*t* 分布は**スチューデントの *t* 分布(Student's *t*-distribution)**とも言い[7]，標本の自由度が少ない場合に有用な分布である[8].

自由度 N の *t* 分布の PDF は，次式で与えられる(柴田 1996, p. 89; 小針 1973, p. 181).

$$W_T(t,N) = \frac{1}{\sqrt{\pi N}} \frac{\Gamma((N+1)/2)}{\Gamma(N/2)} \left(1+\frac{t^2}{N}\right)^{-\frac{N+1}{2}} = \frac{1}{\sqrt{N} B\left(\frac{1}{2},\frac{N}{2}\right)} \left(1+\frac{t^2}{N}\right)^{-\frac{N+1}{2}} \tag{1.22}$$

これら二つの表現が一致することは，$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ より，$B\left(\frac{1}{2},\frac{N}{2}\right) = \frac{\Gamma(1/2)\Gamma(N/2)}{\Gamma((1+N)/2)} = \sqrt{\pi}\,\frac{\Gamma(N/2)}{\Gamma((1+N)/2)}$ となることから容易に見て取れる．

また，　*t*-分布と *F* 分布は，*t* が自由度 *N* の *t*-分布に従うならば，t^2=*F* とおくとき，*F* は自由度(1,*N*)の *F* 分布に従う，という強い関係を持っている[9]．この性質のために，例えば相関係数の有意水準の判定は，*t*-分布を用いても *F*-分布を用いても行なうことができ，同じ結果が得られる．

---

[7] Student は，この分布を研究した W. L. Gossett のペンネームである(Thiebaux, p. 155).
[8] 標本の自由度が多い場合には，t-分布でなく，正規分布が使われる．
[9]証明は，小針 (1973), p. 183.

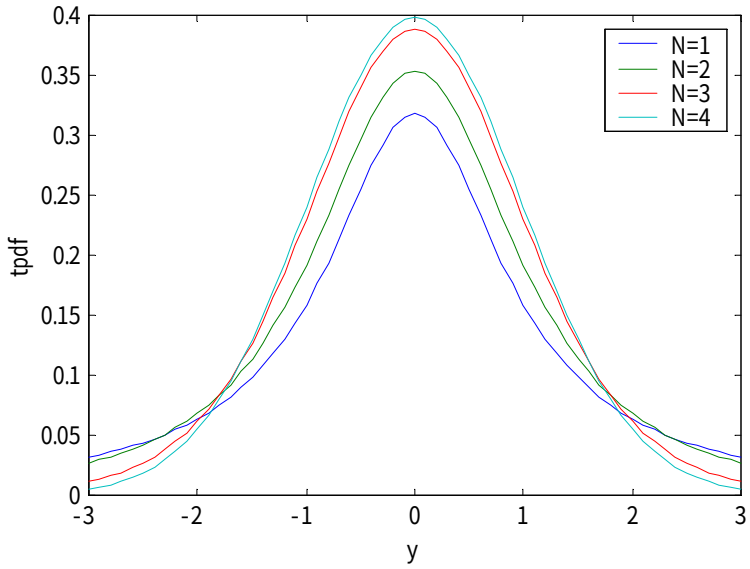

図 1.4. 自由度 1,2,10,100 の *t* 分布の PDF.

1.8. 推定

推定(estimation)とは，限られた標本のデータをもとに，母集団についての情報を得る操作である．推定された母集団パラメータの値を，**推定量(estimator)**と呼ぶ．一つの値を得る推定を点推定(point estimation)という．例えば，標本平均は点推定である．一方，例えば 95%の確率で真の値を含む範囲を推定しようということもある．このような真の値が存在するであろう区間を推定することを，区間推定(interval estimation)という．例えば，真の値が 95%の確率で，区間 (θ_l,θ_u) に存在すると推定するのであれば（添え字 *l*,*u* は lower, upper を意味する），95%を 0.95 と示して**信頼度(confidence level)**あるいは信頼係数(confidence coefficient)[10]，区間を**信頼区間(confidence interval)**，区間の上限と下限である θ_l と θ_u を**信頼限界(confidence limits)**と呼ぶ．なお，大気海洋では，点推定は通常は統計的な推定であると意識されることは少なく，統計的推定とはほとんどの場合区間推定を意味する．

1.8.1. 一致性と不偏性

母集団のパラメータの推定値が，推定に用いる実現値の数を増やした場合に，真のパラメータの値に漸近する場合，その推定値および推定方法は**一致性(consistency)**を持つという．実現値とは例えば年でサンプルされる時系列の平均を推定するのであれば，使えるデータの年数が実現値の数である．

一方，例えば多くの気象観測データは 100 年間しか存在しないというように，そもそも実現値

[10] Emery and Thomson 1997 p. 216 には，confidence level と significance level とは同じ意味だと書かれているが，これは正しくないと思う．95%=1- αと表した， αが有意水準であろう．ただし区間推定では有意水準という言葉は使われない．推定値が信頼区間の中に入るという帰無仮説を立てて，それが棄却されるかどうかを調べる場合には，有意水準という用語を使うのがふさわしいだろう．

の数は限定されていることを前提にしなければならない場合も多い．一致性を持つとは，実現値の数を無限にすれば，推定が妥当であるということを意味するだけで，実現値の数が有限である場合の情報を与えるものではない．ここで，仮にある母集団からある個数のデータ（例，100 年の時系列）の組・セットが，多数得られるという状況を考えよう．つまり，100 年間の時系列が，10 セットでも，100 セットでも存在するとする．この場合，100 年間の時系列についての統計量のある推定方法を，複数のセットについて平均することで，その推定方法が良いかどうかを判断できるだろう．有限個の実現値のセットに基づく，母集団のパラメータの推定値が，セットの数を増やした場合に，真のパラメータの値に漸近する場合，その推定値は**不偏推定 (unbiased estimator)**であるという．そうでない場合にはその推定値は，**偏りを持つ（biased）**という．

1.8.2. 不偏分散

推定に偏りが生ずる例の代表は分散の推定である．そこで相関で登場した不偏分散を導出しておこう．母平均が既知であれば，標本分散は

$$s^2 = \frac{1}{N}\sum_{i=i}^{N}(x_i - \mu)^2$$

で得ることができる．しかし，母平均を標本平均

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

で置きかえると，

$$V^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x}) \tag{1.23}$$

で求めなくてならない. この分散がどのような性質を持つかを検討するために, その期待値(⟨ ⟩で表す)を調べよう.

$$\begin{aligned}
\left\langle V^2 \right\rangle &= \left\langle \frac{1}{N}\sum_{i=1}^{N}(x(i)-\overline{x})^2 \right\rangle \\
&= \left\langle \frac{1}{N}\sum_{i=1}^{N}\left[(x-\mu)-(\overline{x}-\mu)\right]^2 \right\rangle \\
&= \left\langle \frac{1}{N}\sum_{i=1}^{N}(x(i)-\mu)^2 \right\rangle - \frac{2}{N}\left\langle \sum_{i=1}^{N}(x(i)-\mu)(\overline{x}-\mu) \right\rangle + \frac{1}{N}\sum_{i=1}^{N}\left\langle (\overline{x}-\mu)^2 \right\rangle \\
&= \sigma^2 - \left\langle (\overline{x}-\mu)^2 \right\rangle
\end{aligned}$$

この第二項は，既に大数の定理の一例として示した，平均された正規分布の分散に他ならず，それはσ^2 / N となるので，結局.

$$\left\langle V^2 \right\rangle = \frac{N-1}{N}\sigma^2$$

となるので，V^2 は母分散よりも N/(N-1)だけ小さい方に偏よりがあるということになる．この偏よりの逆数をかけて補正したのが，不偏分散すなわち，偏りのない母分散の推定量は

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x(i)-\bar{x})^2 \tag{1.24}$$

である．この資料では(1.23)を偏分散(biased variance)，(1.24)を**不偏分散(unbiased variance)**と呼んで区別する．なお，偏分散・不偏分散ともに一致性を持っている．

また，相関係数・回帰係数の計算では，分散の計算で偏分散と不偏分散のどちらの流儀を採用しても，その流儀を首尾一貫して使えば同一の結果を得るので，流儀をさほど気にしなくても正しい計算ができる．

1.8.3. 母分散が既知の場合の母平均値の区間推定

母分散のσ^2が既知で，母集団が正規分布するか，標本の個数 N が十分に大きい(30 個以上)であれば，母平均 μ の信頼度 1- αの信頼区間は次式で与えられる．

$$\bar{x} - z(\alpha/2)\frac{\sigma}{\sqrt{N}} < \mu < \bar{x} + z(\alpha/2)\frac{\sigma}{\sqrt{N}} \tag{1.25}$$

また同じ内容を，母平均がこの区間に入る確率 P を用いて，次のように表現できる．

$$P\left(\bar{x} - z(\alpha/2)\frac{\sigma}{\sqrt{N}} < \mu < \bar{x} + z(\alpha/2)\frac{\sigma}{\sqrt{N}}\right) = 1-\alpha \tag{1.26}$$

ここで， $z(\alpha/2)$ は，標準正規分布の PDF をW_N として，

$$\int_{-\infty}^{-z_{\alpha/2}} W_N(x)dx = \int_{+\infty}^{z_{-\alpha/2}} W_N(x)dx = \alpha/2, \ \Leftrightarrow \ \int_{-z_{\alpha/2}}^{z_{\alpha/2}} W_N(z)dz = 1-\alpha \tag{1.27}$$

で定義される． CDF を F_N として

$$F_N(-z(\alpha/2)) = \alpha/2,\ F_N(z(\alpha/2)) = 1-\alpha/2 \tag{1.28}$$

とも表現できる．

(1.27)(1.28)式の意味は PDF，CDF の扱いに慣れていないと分かりづらいので，図 2.5 を示して説明しよう．まず，PDF で考えるると，中央部が高く端が低い．PDF は z=0 に対象な分布をしているので，ある特定の z1 よりも大きい z の実現値を取る確率と，-z1 よりも小さい z の実現値を取る確率は同じである．この確率は，PDF の $(-\infty, z_1)$ または (z_1, ∞) の積分で表される．この時に両者の確率を α/2（中央の部分に入る確率が 1- α）として，その場合の特定の z の値を $z_{\alpha/2}$ で表すことにする．一方，CDF で考えられば，$F_N(-z(\alpha/2))$ は PDF の左端の面積であるから$\alpha/2$ となるし， $F_N(z(\alpha/2))$ は PDF の左端と中央の面積の和であるので$1-\alpha/2$ となる．

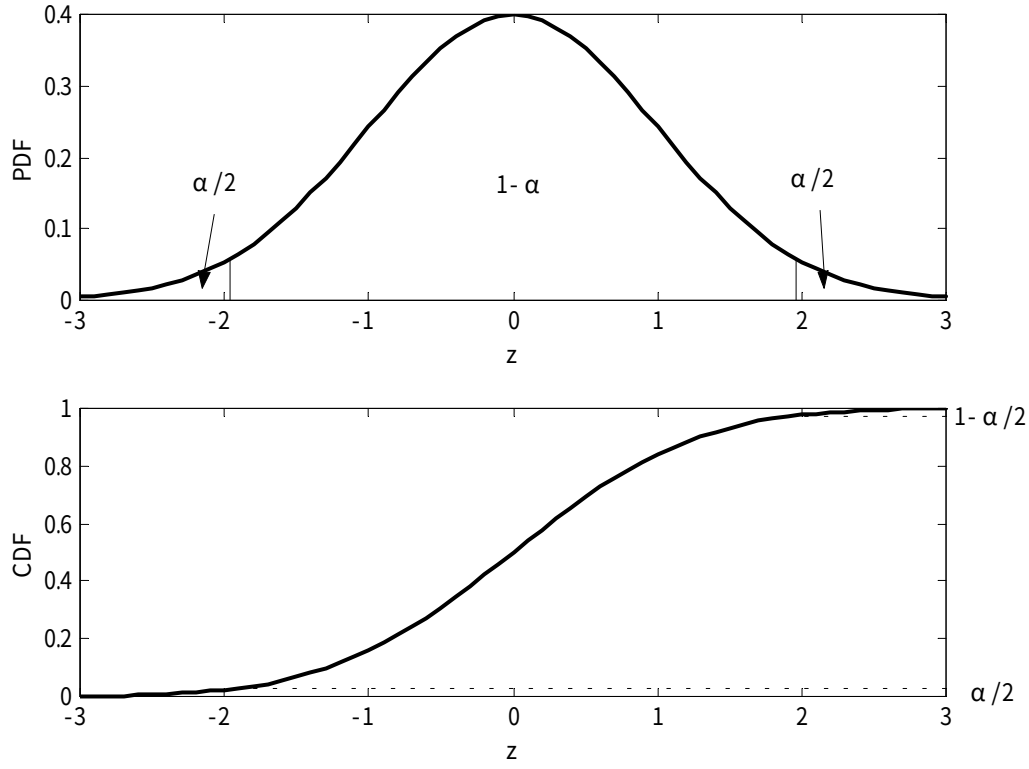

図 1.5. 標準正規分布の PDF（上）と CDF（下）についての，$\alpha = 0.05$ の場合の(1.27)(1.28)の図解.

実際の応用では，適当に決めたα に対して $z_{\alpha/2}$ を求めることが必要である．正規分布の ICDF を $\tilde{F}_N$ で表せば,

$$z(\alpha/2) \equiv \tilde{F}_N(1-\alpha/2) = -\tilde{F}_N(\alpha/2) \tag{1.29}$$

となるので，Matlab では **norminv** を Octave では **normal_inv** を用いて $z_{\alpha/2}$ を計算することができる．なお，どの程度の α を使えばよいかというのは分野によってもちがうかもしれない．気候変動研究では，推定・検定では$\alpha = 0.05$ を使うことが多く，この値であれば α が大きすぎる（つまり推定が甘すぎる）という非難を受けることはないだろう．これよりも小さい値であれば，標準的な研究よりも厳密であるという印象を与えることができる．また，$\alpha = 0.10$ は大き目であるけれど，許容される範囲だろう．ただし多くの応用で本来問題にすべきなのは，αをどう決めるかではなく，母集団は正規分布するか標本の数が十分に多いかの，どちらかの前提が満たされているかであろう.

(1.26)の**証明：**すでに説明したように，$\overline{x} = \sum_{i=1}^{N} x(i)/N$ は，正規分布 $N(\mu, \sigma^2/N)$ に従うので，

$z = \dfrac{\overline{x} - \mu}{\sigma/\sqrt{N}}$ は標準正規分布 $N(0,1)$に従う．よって，　$P\left(-z(\alpha/2) < \dfrac{\mu - \overline{x}}{\sigma/\sqrt{N}} < z(\alpha/2)\right) = 1-\alpha$

が定義から成り立ち，式を整理して(1.26)を得る.

1.8.4. 母分散が未知の場合の母平均値の区間推定

上の母分散が既知であるという状況は実際にはありそうにない．そこで，これらの場合には，自由度 N-1 の t-分布を用いて，次のように母平均の区間推定を行なう．

$$P\left(\bar{x}-t(\alpha/2,N-1)\frac{s}{\sqrt{N}}<\mu<\bar{x}+t(\alpha/2,N-1)\frac{s}{\sqrt{N}}\right)=1-\alpha \tag{1.30}$$

ここで S^2 は不偏分散であり，また $t(\alpha/2,\nu)$ は t-分布の ICDF を $\tilde{F}_t$ として，$t(\alpha/2,\nu)=\tilde{F}_t(1-\alpha/2,\nu)$ で与えられる[11]．なお，この場合も母集団が正規分布するという仮定は用いている．

1.8.5. 母分散の区間推定

母平均が未知の場合の[12]，母分散のσ^2 の信頼率 αの信頼区間は

$$P\left(\frac{S(N-1)}{\tilde{F}_\chi\left(\frac{\alpha}{2},N-1\right)}<\sigma^2<\frac{S(N-1)}{\tilde{F}_\chi\left(1-\frac{\alpha}{2},N-1\right)}\right)=1-\alpha \tag{1.31}$$

で与えられる．ただし，

(1.31)の証明：統計量 $\chi^2=S(N-1)/\sigma^2=\sum_{i=1}^{N}(x(i)-\bar{x})^2/\sigma^2$ が自由度 N-1 の χ^2 分布に従うので，

$$P\left(\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)<\chi^2<\chi^2_{n-1}\left(\frac{\alpha}{2}\right)\right)=1-\alpha$$

となる．ここで括弧内各項の逆数を取り，$\chi^2=S(N-1)/\sigma^2$ を代入することで上式を得る．

1.8.6. 相関係数の区間推定

相関係数 r を，$z=\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$ というフィッシャーの z 変換(Fisher's z-transform)を用いると，z は $N\left(\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right),\frac{1}{n-3}\right)$ である正規分布に漸近的に収束することが知られている．これから，z の有意水準 αでの推定区間は，

$$\hat{z}-\frac{Z_{\alpha/2}}{\sigma_N}<z<\hat{z}+\frac{Z_{\alpha/2}}{\sigma_N},\ \sigma_N=\frac{1}{\sqrt{n-3}} \tag{1.32}$$

また，$r=\tanh(z)$ なので，相関係数自体の信頼区間は，

$$\tanh(\hat{z}-Z_{\alpha/2}/\sigma_N)<r<\tanh(\hat{z}+Z_{\alpha/2}/\sigma_N) \tag{1.33}$$

---

[11] 証明は柴田 1996 p. 110-114 を参照.
[12] 母平均が既知の場合については柴田 1996 p.177 を参照.

で与えられる．なお，ここでハットは観測値を示している．ただし上の表現において，サンプル数でなく自由度を用いるにはどうしたら良いかは不明．なお，Emery and Thomson (1997)の p. 253 に，は Storch and Zwiers (1999)の p.148 にも示されている．

1.9. 検定

統計**検定**(statistical **test**)とは，ある仮説 H_0 が**棄却(reject)**できるかどうかを，統計的に調査することである．この仮説を**帰無仮説(null hypothesis)**とよび，そのためこの種の検定を **null-hypothesis** test と呼ぶこともある．検定では通常，あるの**有意水準(significance level, level of significance)**α(信頼度 1-α)を設定し，その有意水準に照らして帰無仮説を棄却できるかどうかを調べる．有意水準はまた危険率とも呼ばれる．例えば，時系列１と時系列２とが相関が無いという帰無仮説を立て，この仮説が有意水準 5%で棄却されるのであれば，時系列１と時系列２との間には実際に相関がある確率は 95%以上であり，実際には相関がない確率は 5%以下である．帰無仮説が棄却できない場合は，実際には相関があるかもしれないし，無いかもしれず，さほど有効な情報は得られず，仮説 H_0 は無に帰すこととなる．このために，帰無仮説との呼び名がある．

実際の検定に当たっては，測定されている値をそれぞれの検定で用いられる特定の式で統計検定量と呼ばれる値をまず計算する．この統計検定量が，t-分布なり，F-分布なりに従うことを利用して，検定を行う．

なお，帰無仮説検定において，有意となる最小の有意水準を **P 値(P-value)**と呼び，これを用いる場合もある．例えば，相関係数の場合であれば，P 値は無相関の帰無仮説が正しい場合に，観測された相関係数よりも絶対値がより大きな相関係数が得られる確率である．帰無仮説検定においては，検定に先立って有意水準を設定することが理想とされているけれど，5%有意水準と 10%有意水準の両方で検定して，5%で有意になるならそれを報告し，10%で有意になるならそれを報告するということが実際には行なわれ勝ちである．それよりも，P 値を報告する方が，より詳しい情報であり恣意性も入らない．

1.9.1. 第一種の誤り・第二種の誤り

検定には，以下の第一種と第二種の誤りがある．

第一種の誤り(error of the first kind, type I error)とは，帰無仮説 H_0 が正しいにもかかわらず，これを誤りとして棄却することである．この危険率は有意水準 αである．

第二種の誤り(error of the second kind, type II error)とは，帰無仮説 H_0 が誤っているにもかかわらず，これを正しいとして採用すること．この確率を βで表す．この誤っている H_0 を正しく棄却する確率，1-βを検出力(power of the statistical test) (Emery and Thomson, 1998, p. 249)と呼ぶ．

		帰無仮説	
		正	誤
判断	正		II β
	誤	I α	

1.9.2. 分散比の検定

分散比の検定は，次の平均の差の検定にも使われるので，その前に解説しておこう．正規分布するデーター$x(1), x(2),\ldots, x(N_x)$と $y(1), y(2), \ldots, y(N_y)$ とがあり，その不偏分散推定値 S_x^2, S_y^2 の比 S_x^2 / S_y^2 は自由度 N_x-1, N_y-1 の F 分布に従うので，

$$S_x^2 / S_y^2 < \tilde{F}_F(\alpha / 2, N_x, N_y) \text{ または } \tilde{F}_F(1-\alpha / 2, N_x, N_y) < S_x^2 / S_y^2 \tag{1.34}$$

であれば，帰無仮説 $\sigma_x^2 = \sigma_y^2$ は棄却される．

1.9.3. 平均の差の検定

正規分布するデーター$x(1), x(2),\ldots, x(N_x)$と $y(1), y(2), \ldots, y(N_y)$ とがあって，両者の間に有意な平均の差があるかどうかを調べるには，まず，分散比の検定を行なって，分散に有意といえるほどの差がなければ，等分散を仮定する平均の差の検定を行なう．この検定では，帰無仮説 $\mu_x = \mu_y$ の下では，

$$T = \frac{\overline{x} - \overline{y}}{\sqrt{\dfrac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{n_x + n_y - 2}}\sqrt{\dfrac{1}{n_x} + \dfrac{1}{n_y}}} \tag{1.35}$$

が自由度 $Nx+Ny+1$ の t-分布に従うことを利用して，T の実現値の絶対値が $T = \tilde{F}_t(1-\alpha / 2,\ N_x + N_y - 2)$ を上回るかどうかを調べる．もし上回れば，帰無仮説を棄却する．

等分散を仮定する平均の差の検定は，等分散でなくとも $N_x \approx N_y$ の場合には，ほぼ正しい結果を与えることが知られている．

もし分散に有意とみなさざるを得ないほどの，大きな差があるのならば，ウェルチ（Welch）の検定を行う．この検定では帰無仮説の下では，

$$T = \frac{\overline{x} - \overline{y}}{\sqrt{\dfrac{S_x^2}{n_x} + \dfrac{S_y^2}{n_y}}} \tag{1.36}$$

が自由度 k の t-分布に近似的に従う．ここで，k は

$$\frac{1}{k}=\frac{c^2}{n_x-1}+\frac{(1-c^2)}{n_2-1},\ c=\frac{\frac{S_x^2}{n_x}}{\frac{S_x^2}{n_x}+\frac{S_y^2}{n_y}}$$

で与えられる．なお，(1.36)は近似的にしか t-分布に従わないので，ウェルチの検定よりも等分散を仮定する平均の差の検定が好まれる．目安として永田(1996)は，分散が倍以上違って，かつサンプルの数が倍以上異なるときにはウェルチの検定を用いるべきだと述べている．

1.9.4. 無相関の検定

それぞれ正規分布に従う N の標本を選ぶ．両者に相関がなければ，検定統計量 $T(r)=\frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$ は，自由度(Degree of Freedom) $N-2$ の t 分布に従うことが知られている．従って，t-分布と F-分布との関係から，$F=\frac{r^2(N-2)}{1-r^2}$ は自由度(1,N-2)の F 分布に従う[13].

まず上の関係式のうち t-分布を使うには, 自由度 N-2 の 1-α/2 信頼度の $T=\tilde{F}_t(1-\alpha/2,\ N-2)$ に対して，$r=\frac{T}{\sqrt{N-2+T^2}}$ が同じ信頼度αでの相関係数の最小値を与え，相関係数の絶対値がこれよりも大きいのであれば有意であると言える．また, 同じことだが, $T=\tilde{F}_t(1-\alpha/2,N-2)$ と $T(r)$ の実現値を比較して有意性を判定してもよい．

F-分布を利用するのも同様に，F の実現値と $F=\tilde{F}_F(1-\alpha,1,N-2)$ とを比較する．t-分布では 1-α/2 を使い，F-分布では 1-αを使うのは，相関係数の絶対値が大きい場合に T の実現値は正負に大きくなるのに大して，F の実現値は正に大きくしかならないためである．信頼度αでの相関係数の最小値は，$r=\sqrt{\frac{F}{N-2+F}}$ である．

演習問題 1.2. 相関係数が有意かどうかをテストするには，上の t-分布と F-分布を用いる方法のほかに，既に説明した相関係数の区間推定を用いる方法，およびモンテカルロ法を用いる方法がある．これらの 4 通りの方法で，標本数が 4〜20 の場合について 5%有意水準で有意となる相関係数を求め，t-分布と F-分布を用いる方法は正確に一致することを，他の方法もほぼ同じ結果を与えることを確認せよ．

---

[13] 証明は簡単ではないが，この導出は和書では，丸山(1958)に示されている．
丸山儀四郎, 1958: 確率および統計 (基礎数学講座) / 秋月康夫[ほか]編, 共立出版, p. 161.

1.9.5. 統計的に有意でも物理的に無意味な場合

帰無仮説の検定には実は様々な批判が加えられている．気象・海洋の分野ではそういった批判も遅れており, 批判を踏まえた上での適切な利用も進んできたとは言えない. しかし, 最近 Nicholls (2001)では帰無仮説であることを強く主張しており, 今後は帰無仮説の有効性にも従来よりも批判的に正しく扱う方向に進むことが期待される. そこで, 帰無仮説にひそむ問題を説明しておこう．

例えばエルニーニョがある地点の気温と関係があるかという問いを，立てたとしよう．もしエルニーニョが十分に長い期間現在のように地球の大気大循環に影響を及ぼしていたのであれば，地球上の全ての地点に大なり小なり，たとえ非常に微小であっても，影響を及ぼしていたであろうと考えられる．したがって，データが十分にありさえすれば，有意な相関係数が得られるであろう．そうであるなら，統計的に有意であるかどうかは，単にデータが十分に長いかどうかを意味するものでしかない．しかし，予測や解釈に有効であるためには，ある程度強い関係でなくてはならない．すなわち，相関係数の絶対値がある程度大きいことが必要である．

自由度が大きい場合には，どの程度の相関係数で有意となるのかを具体的に調べてみよう．例えば，自由度が 10 の場合に，相関が 5%有意水準で有意となるためには，相関係数は 0.58 以上でなくてはならない．一方自由度が 50, 100 なら有意となるのに必要な相関係数は，それぞれ 0.28, 0.20 である．しかし，相関係数が 0.28 ということは，その関係で説明されるエネルギーは 8%に過ぎない．この場合，統計的に有意であっても，物理的には無意味である．この問題は，統計学者の間ではよく知られており，例えば永田(1996)では，無相関の検定は，有意でありさえすれば意味があるという大誤解を招くので，初学者には教えない方が良いという大先生もいることを紹介している．

ではどの程度の相関係数の値であれば，物理的に意味のある相関係数だろうか？Nicholls (2001)は相関係数が 0.6 以上，永田(1996)は相関係数が 0.5〜0.6 以上であれば，意味のある値と述べている．相関係数が 0.5, 0.6, 0.7 で，前に述べたように説明されるエネルギーの割合が約 1/4, 1/3, 1/2 であるから，0.5〜0.6 というのは分かりやすい目安だろう．相関係数が 0.3 程度であっても統計的に有意だと言い立てる論文があるが，以上の理由から実際的な意味はほとんどなく，そういった論文は統計理論に振りまわされていると言える．もし相関係数が 0.3 程度であれば，より強い相関が得られるように季節依存性や時間スケール依存性を調査するか，その路線はあきらめるべきだろう．

また，Nicholls (2001)は帰無仮説検定ではなく，相関係数の区間推定を行う方が良いと主張し，Gardner and Altman (1989)に，相関係数，平均値の差，その他もろもろの信頼区間の推定法が示されていることを述べている．なお，Nicholls (2001)は，２次元空間に相関係数をマップするような目的では，帰無仮説検定を行い有意水準のコンターを描く方が現実的であることを認めている．ただしその場合でも，特に注目する領域において相関係数の区間推定をするべきであると主張している．

Nicholls (2001)の帰無仮説が無意味であるという主張は，頭に入れておくべき重要性を持っている. 相関係数でも単に真の相関がゼロでないかどうかを議論するのではなく, 区間推定によって, 真の相関がどの範囲に入るかを示す方が, より適切な情報が得られる. ただしここ数年のうちに, 帰無仮説検定が大気・海洋の研究からなくなることはないだろう．たしかに，統計量の区間推定ができるのであれば，その方が優れている場合は多く，その場合に Nicholls (2001)の論文を引用し

て主張を強めるのもお勧めである．しかし研究文化には一定の慣性があるので，相関などの区間推定の利用は広がるであろうけれど，例えば２・３年のうちに有意性検定を上回って利用されるとは思われない．なお，Monte-Carlo 法を利用する場合などは，有意性検定は通常簡単にプログラムできけれど，統計量の区間推定は難しいという問題もある．

1.10. 参考文献

Emery, W. J., and Thomson, R. E., 1997: Data analysis methods in physical oceanography. Pergamon press, pp. 634.

Gardner,M. J., and D. G. Altman, 1989: Statistics with confidence intervals and statistical guideline. British Medical Journal, pp140.

小針　あき宏，1973: 確率・統計入門, 岩波書店, pp. 300.　　# 小針(1973)は急逝した著者の遺稿を４名の友人がまとめたものである．歴史に残る快作である．こんなこと書いたら同僚になんていわれるだろう，などということを考えてしまう凡人には書けない．今日でこそ，こういったセンスの本もないわけではないが, これを 1970 年代前半に成し遂げたのは, 凄い．1970 年頃の教科書は今日見劣りがするものも少なくないが，この本は例外である．

永田　靖，1996: 統計的方法のしくみ ―正しく理解するための 30 の急所 ―．日科技連，p. 238. ISBN4-8171-0294-2

Nicholls, N. 2001: The insignificance of significance testing. *Bull, Am. Met. Soc*., 82, 981-985.

柴田 文明, 1996: 理工系の基礎数学７：確率・統計，岩波書店, pp.217.

Trenberth, K., E., 1984: Some effects of finite sample size and persistence on meteorological statistics. *Mon. Wea. Rev.*, **112** (Dec), 2359-2368.

統計学の用語集が　http://www.cas.lancs.ac.uk/glossary_v1.1/Alphabet.html　にある．

von Stroch, H. and F. W. Zwiers, 1999: Statistical analysis in climate research. Cambridge University Press, pp. 483 （ISBN: 0 521 45071 3 ）