付録１　標本平均の期待値と分散

変量 Xの母平均が μ，母分散が σ^2のとき，標本平均$\overline{X}$の期待値は μ，分散は σ^2/n になる．この性質は以下のように，**期待値と分散の加法性**を使って示せる．

期待値については，以下の加法性が成立する．

$$E(X+Y)=E(X)+E(Y)$$

分散については，変数 Xと Yが独立な場合は加法性が成立する【⇒下記補足】.

$$V(X+Y)=V(X)+V(Y)$$

上記の期待値と分散の加法性から，$X_1, X_2, \ldots, X_n$が独立な場合，標本平均 $\bar{X}$ の期待値と分散について，次の関係が導ける．

$$\mathrm{E}(\bar{X})=\mathrm{E}\left(\frac{X_1+X_2+\cdots+X_n}{n}\right)=\frac{1}{n}\{\mathrm{E}(X_1)+\mathrm{E}(X_2)+\cdots+\mathrm{E}(X_n)\}$$

$$=\frac{1}{n}(\mu+\mu+\cdots+\mu)=\mu$$

$$\mathrm{V}(\bar{X})=\mathrm{V}\left(\frac{X_1+X_2+\cdots+X_n}{n}\right)=\frac{1}{n^2}\{\mathrm{V}(X_1)+\mathrm{V}(X_2)+\cdots+\mathrm{V}(X_n)\}$$

$$=\frac{1}{n^2}(\sigma^2+\sigma^2+\cdots+\sigma^2)=\frac{\sigma^2}{n}$$

【補足】分散の和の期待値（『独習統計学 24 講』より転載）

変量 x，y が独立のとき，$x+y$ の分散はそれぞれの分散の和になる．

$$V(x+y) = V(x)+V(y)$$

証明はやさしいので，以下を見る前に自分でも考えてみよう．なお，μ_x, μ_y は変量 x，y の期待値とする．

$$\begin{aligned}V(x+y) &= \mathrm{E}\{((x+y)-(\mu_x+\mu_y))^2\}\\ &=\mathrm{E}\{((x-\mu_x)+(y-\mu_y))^2\}\\ &=\mathrm{E}\{(x-\mu_y)^2+2(x-\mu_x)(y-\mu_y)+(y-\mu_y)^2\}\\ &=\mathrm{E}\{(x-\mu_y)^2\}+2\cdot\mathrm{E}\{(x-\mu_x)(y-\mu_y)\}+\mathrm{E}\{(y-\mu_y)^2\}\\ &=V(x)+2\cdot\mathrm{E}\{(x-\mu_x)(y-\mu_y)\}+V(y)\end{aligned}$$

ところで x，y は独立なので

$$\mathrm{E}\{(x-\mu_x)(y-\mu_y)\}=\mathrm{E}(x-\mu_x)\times\mathrm{E}(y-\mu_y)=0\times0=0$$

よって

$$V(x+y)=V(x)+V(y)$$

付録 2　2 次元の正規分布と相関係数 ρ の関係（中級）

本文では，相関係数を基準化した共分散としてとらえたが，相関係数は 2 変量の正規分布の母数としても定義できる．連続変量が 1 個の場合，その分布は標本であればヒストグラム，母集団であれば確率密度関数で表すことができた．例えば，第 3 講の図 3.1 左は身長の度数分布，図 3.1 右は確率密度関数の例で，1 変量の場合，確率密度関数は曲線として表せた．変量が 2 個の場合は，変量の値の組(X_k, Y_k)に対して z 軸にその頻度を描けば度数分布や確率密度関数を面として表すことができる．

正規分布の場合の実際の確率密度関数の形を見ておこう．単変量の場合，平均 μ，分散 σ^2 の正規分布 $N(\mu, \sigma^2)$の確率密度関数は，

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2}) \qquad \text{(A2.1)}$$

だった．2 変量の場合，相関がなく，

$$x\sim N(\mu_x,{\sigma_x}^2), y\sim N\left(\mu_y,{\sigma_y}^2\right)$$

のとき，**同時確率密度関数** $f(x, y)$は x，y それぞれの確率密度関数の積で表せる．

$$f(x, y) = f_x(x)\times f_y(y)$$

$$=\frac{1}{\sqrt{2\pi}\sigma_x}\exp(-\frac{(x-\mu_x)^2}{2{\sigma_x}^2})\times\frac{1}{\sqrt{2\pi}\sigma_y}\exp(-\frac{(x-\mu_y)^2}{2{\sigma_y}^2})=\frac{1}{2\pi\sigma_x\sigma_y}\exp(-\frac{(x-\mu_x)^2}{2{\sigma_x}^2}-\frac{(y-\mu_y)^2}{2{\sigma_y}^2})$$

それに対して，相関がある場合は，同時確率密度関数は以下のようになる：

$$f(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp(-\frac{1}{2(1-\rho^2)}[\frac{(x-\mu_x)^2}{{\sigma_x}^2}-\frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}+\frac{(y-\mu_y)^2}{{\sigma_y}^2}]) \qquad \text{(A2.2)}$$

この式中の ρ が相関係数であり，ρ=0 のときは相関がない場合の確率密度関数と一致する．

式だけでは分布がどうなっているかわかりにくいので，$f(x, y)$ をグラフに描いてみると，図 A2.1 のように 1 峰性の曲面になる．図 A2.1 左は相関がまったくない場合，右図は相関がある場合（ρ=0.6）である．このように，相関係数 ρ は 2 次元の正規分布で，2 つの変量がどれくらい関連が強いかを表す母数としても定義できる．

【補足】同時確率密度関数と確率の計算

連続変量の確率の計算について考える．1 変量の場合，ヒストグラムの下の面積が 1 になるように図を描けば，標本の大きさが無限大になった場合として確率密度関数が自然に導入できた. 2 変量の場合は, 度数分布曲面の下の体積が 1 になるようにすれば, 自然に確率密度曲面が得られる．同時確率密度関数はこの曲面を数式で表したもので, 2 変量のある範囲に対する曲面下の体積が，2 変量がその範囲にある確率を表す.

以上の関係を式で書くと，1 変数の場合，確率密度関数を $f(x)$ とすると

$$\Pr(a \le x \le b) = \int_a^b f(x)dx$$

だったが（⇔曲線下の面積の計算），2 変数の分布の場合は，ある領域の確率は同時確率密度関数を二重に積分することで計算できる（⇔曲面下の体積の計算）.

$$\Pr(a \le x \le b, c \le y \le d) = \int_c^d \int_a^b f(x,y)dxdy$$

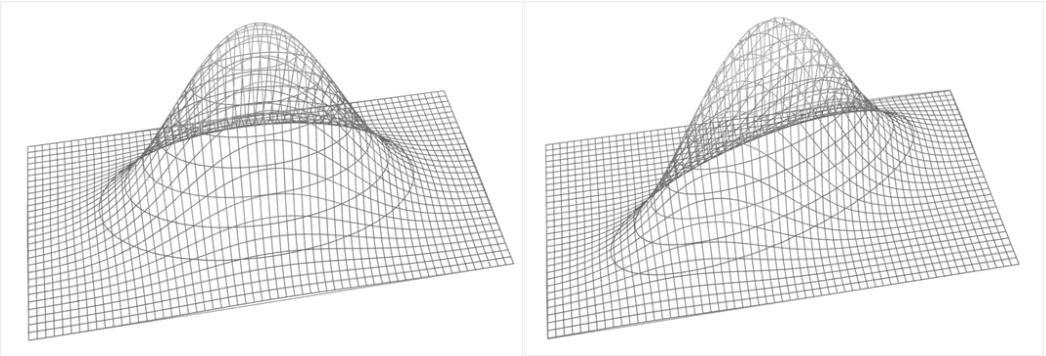

図 A2.1　2 変量の正規分布の確率密度関数（左：$\rho = 0$，右：$\rho = 0.6$）

付録３　*F* 分布：２つの分散の比の分布

２つの群 A，B の平均値が等しいかどうかを調べたいことがよくあるように，２つの群の分散を比較したいこともよくある．A 群，B 群とも標準正規分布をしており，そこから抽出した標本の大きさが n，m，(6.2)式で定義した χ^2 をそれぞれ χ_A^2，χ_B^2 としたとき，

$$F_0=(\chi_A^2/n)/(\chi_B^2/m)$$

についてまず考えてみよう．A群の観測値を x_1，x_2，x_3，…，　x_n，B 群の観測値を y_1，y_2，y_3，…，y_m とすれば，F は

$$F_0=\frac{\frac{x_1^2+x_2^2+\cdots+x_n^2}{n}}{\frac{y_1^2+y_2^2+\cdots+y_m^2}{m}} \qquad \text{(A3.1)}$$

と書き下せる．母平均=0 なので，x_k^2，y_k^2 は偏差の二乗でもある．また，両群の母分散は等しいので，分母と分子の比 F_0 は 1 に近いことが予想されるが，単純に特定の分布にはならず，n，m の値によって形が少しずつ異なり，

自由度 *n*，*m* の *F* 分布

と呼ばれる，図 A3.1 のような右下がりまたは単峰性の分布になる．

A 群，B 群が標準正規分布でなく一般の正規分布$N(\mu_x,\sigma_x^2)$，$N(\mu_y,\sigma_y^2)$に従う場合は，基準化

$$z_k=\frac{x_k-\mu_x}{\sigma_x},\quad w_k=\frac{y_k-\mu_y}{\sigma_y}$$

を行えば z_k，w_k は標準正規分布に従うので，z_k，w_k を使って F_0 の式(A3.1)の形は書ける．しかし，母平均 (μ_x,μ_y) も母分散 (σ_x,σ_y) もふつうは未知なので，残念ながらこのままでは F_0 の値は計算できない（式は書けるが，未知の母数が 4 個もある！）．

そこで，母分散の推定のときと同じように，χ_A^2，χ_B^2 を S_{xx}/σ_x^2，S_{yy}/σ_y^2 で置き換えてみよう（S_{xx}, S_{yy} は測定値の偏差の二乗和）．さらに，S_{xx}/σ_x^2，S_{yy}/σ_y^2 は，それぞれ自由度は$(n-1)$，$(m-1)$の χ^2 分布をするので**【⇒第 6 講 e 節】**，$(n-1)$，$(m-1)$で割ってその比をとると，以下のような統計量が得られる．

$$F=\left(\frac{S_{xx}/\sigma_x^2}{(n-1)}\right)/\left(\frac{S_{yy}/\sigma_y^2}{(m-1)}\right)=\left(\frac{S_{xx}/(n-1)}{\sigma_x^2}\right)/\left(\frac{S_{yy}/(m-1)}{\sigma_y^2}\right)$$

$$=\left(\frac{\hat{\sigma}_x^2}{\sigma_x^2}\right)/\left(\frac{\hat{\sigma}_y^2}{\sigma_y^2}\right) \qquad \text{(A3.2)}$$

ここで，$\hat{\sigma}_x$，$\hat{\sigma}_y{}^2$は第 6 講の(6.4)式の定義に基づく不偏分散である．これで，未知の母数は（σ_x，σ_y）の 2 個に減った！　このとき，

F は自由度$(n-1)$，$(m-1)$の F 分布に従う

ことが知られている**【⇒竹内(1963)】**．そこで，自由度$(n-1)$，$(m-1)$の F 分布の下位 2.5%点と上位 2.5%点の値を，それぞれ

$$F_{n-1,m-1}(0.025),\quad F_{n-1,m-1}(0.975)$$

で表すと，$F=\left(\frac{\hat{\sigma}_x{}^2}{\sigma_x{}^2}\right)\Big/\left(\frac{\hat{\sigma}_y{}^2}{\sigma_y{}^2}\right)$がこの間の値である確率は 95%なので

$$\Pr\{F_{n-1,m-1}(0.025)\le\left(\frac{\hat{\sigma}_x{}^2}{\sigma_x{}^2}\right)\Big/\left(\frac{\hat{\sigma}_y{}^2}{\sigma_y{}^2}\right)\le F_{n-1,m-1}(0.975)\}=0.95$$

という関係が得られる．未知の母数は 2 つあるが，母標準偏差の比 $\frac{\sigma_y}{\sigma_x}$ を未知の母数とみると（未知の母数が実質 1 個になった！），第 6 講で母分散の信頼区間を求めたときと同じ手順により，Pr{ } の中は

$$F_{n-1,m-1}(0.025)\times\left(\frac{\hat{\sigma}_y}{\hat{\sigma}_x}\right)^2\le\left(\frac{\sigma_y}{\sigma_x}\right)^2\le F_{n-1,m-1}(0.975)\times\left(\frac{\hat{\sigma}_y}{\hat{\sigma}_x}\right)^2$$

となる．次にこの式の逆数をとると，$\left(\frac{\sigma_x}{\sigma_y}\right)^2$について解くことができ，

$$\frac{1}{F_{n-1,m-1}(0.975)}\left(\frac{\hat{\sigma}_x}{\hat{\sigma}_y}\right)^2\le\left(\frac{\sigma_x}{\sigma_y}\right)^2\le\frac{1}{F_{n-1,m-1}(0.025)}\left(\frac{\hat{\sigma}_x}{\hat{\sigma}_y}\right)^2 \qquad \text{(A3.3)}$$

となって，分散の比の信頼区間が求まる．

また，両群の分散が等しいかどうか検定を行いたいときは，帰無仮説は

$$\mathrm{H}_0:\sigma_x{}^2=\sigma_y{}^2$$

とすればよいので，(A3.2)式から未知の母数が消えて

$$F=\hat{\sigma}_x{}^2/\hat{\sigma}_y{}^2$$

となるので値を求めることができる．しかも，F は自由度$(n-1)$，$(m-1)$の F 分布に従うことがわかっているので，F を検定統計量とすれば，（帰無仮説が正しいとき）F が下位 2.5%点と上位 2.5%点の間にある確率は 0.95，つまり

$$\Pr(F_{n-1,m-1}(0.025)\le F\le F_{n-1,m-1}(0.975))=0.95$$

なので，下位 2.5%点と上位 2.5%点の外側を棄却域とすれば，2 群の分散が等しいという帰無仮説の検定を行うことができる．

【補足 1】上記の演繹過程では，帰無仮説が両群の分散が等しい，つまり $\sigma_x^2=\sigma_y^2$ の場合，統計量 F の式から未知の母数 σ_x，σ_y が消えて F の値が標本から計算できるようになり，さらに F の分布はわかっていることを利用した.

【補足 2】この検定は, 外れ値 (1 つだけ他の測定値よりかけ離れた測定値) があると，うまく機能しないことがある．そのため，事前に外れ値の確認を行うことが望ましい**【⇒広津，2004】**.

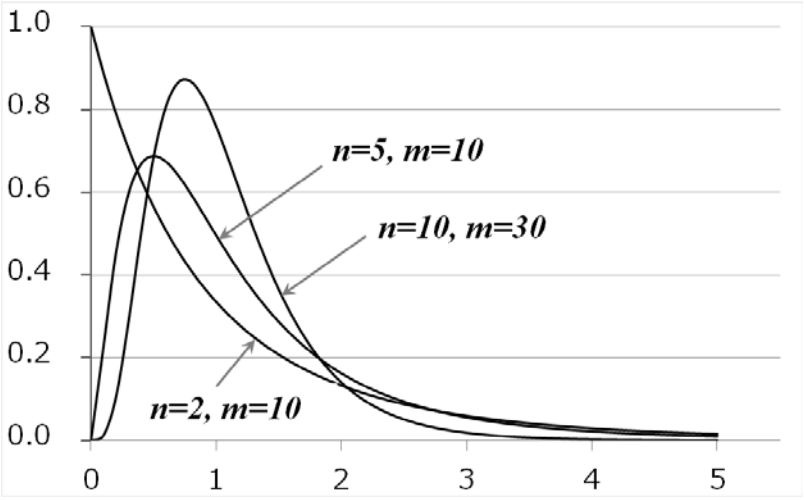

図 A3.1　F分布の例

付録 4　単回帰（中級）

線形回帰モデルは，線形代数（行列とベクトル）を使うと単回帰と重回帰を統一的に，かつ数学的に明快に扱うことができるが，以下では高校数学の範囲内というこの本の方針を考えて，通常の数式による説明を行う．ただし，高校数学の範囲を超える部分は概説にとどめた．さらに詳しい議論を学びたい読者は，**広津（1992）**，**永田・棟近（2001）**，**蓑谷（2004）**，**竹内（1963）**などを参照されたい．

A4.1　回帰係数の点推定値

回帰係数 α，β の点推定値 $\hat{\alpha}$，$\hat{\beta}$ を最小二乗法で求めてみよう．まず，回帰直線

$$Y = \hat{\alpha} + \hat{\beta}X$$

が推定できたとしてみよう．このとき，残差 e_k は

$$e_k = Y_k - (\hat{\alpha} + \hat{\beta}X_k)$$

なので，残差平方和 S_e は以下のように表せる．

$$S_e = \sum_{k=1}^{n} {e_k}^2 = \sum_{k=1}^{n} \left(Y_k - \left(\hat{\alpha} + \hat{\beta} \cdot X_k\right)\right)^2 \qquad \text{(A4.1)}$$

ここで，(X_k, Y_k) は測定値であり定数とみなせるので，未知数は $\hat{\alpha}$，$\hat{\beta}$ の 2 個である．そこで，S_e を $\hat{\alpha}$，$\hat{\beta}$ の関数と見ると，S_e は $\hat{\alpha}$，$\hat{\beta}$ の絶対値が大きくなると値が大きくなるので下に凸である．したがって，残差平方和 S_e が最小となる $\hat{\alpha}$，$\hat{\beta}$ では，それぞれの変数による微分（正確には偏微分と呼ぶ）がゼロである．よって，以下の方程式を満たさなくてはいけない（X_k，Y_k は定数であることに注意して，鉛筆をもって(A4.1)式を微分してみよう）．

$$\frac{\delta}{\delta \hat{\alpha}} S_e = -2 \sum_{k=1}^{n} \left(Y_k - \left(\hat{\alpha} + \hat{\beta} \cdot X_k\right)\right) = 0 \qquad \text{(A4.2)}$$

$$\frac{\delta}{\delta \hat{\beta}} S_e = -2 \sum_{k=1}^{n} X_k \left(Y_k - \left(\hat{\alpha} + \hat{\beta} \cdot X_k\right)\right) = 0 \qquad \text{(A4.3)}$$

さらに，未知数$\hat{\alpha}$，$\hat{\beta}$について整理すると以下の形になる．

$$n\cdot\hat{\alpha}+\left(\sum_{k=1}^{n}X_k\right)\hat{\beta}=\sum_{k=1}^{n}Y_k$$

$$\left(\sum_{k=1}^{n}X_k\right)\hat{\alpha}+\left(\sum_{k=1}^{n}{X_k}^2\right)\hat{\beta}=\sum_{k=1}^{n}X_k\cdot Y_k \qquad \text{(A4.4)}$$

この連立方程式を**正規方程式**と呼ぶ．求める $\hat{\alpha}$，$\hat{\beta}$ については 1 次式なので，あとは中学数学の連立一次方程式の問題である．(A4.4)の解を平方和 S_{xx}，S_{xy} を使って表すと以下のようになる．

$$\hat{\beta}=\frac{S_{xy}}{S_{xx}},\quad \hat{\alpha}=\overline{Y}-\hat{\beta}\overline{X} \qquad \text{(A4.5)}$$

続いて，今求めた回帰係数の推定値 $\hat{\alpha}$，$\hat{\beta}$ を(A4.1)式に代入すれば，残差平方和 S_eの最小値 $S_{e,\min}$を求めることができる．

$$\begin{aligned}S_{e,\min}&=\sum_{k=1}^{n}\left(Y_k-\left(\hat{\alpha}+\hat{\beta}\cdot X_k\right)\right)^2\\&=\sum_{k=1}^{n}\left(Y_k-\left((\overline{Y}-\hat{\beta}\overline{X})+\hat{\beta}\cdot X_k\right)\right)^2=\sum_{k=1}^{n}\left((Y_k-\overline{Y})-\hat{\beta}(X_k-\overline{X})\right)^2\\&=\sum_{k=1}^{n}(Y_k-\overline{Y})^2-2\hat{\beta}\sum_{k=1}^{n}(Y_k-\overline{Y})(X_k-\overline{X})+\hat{\beta}^2\sum_{k=1}^{n}(X_k-\overline{X})^2\\&=S_{yy}-2\hat{\beta}S_{xy}+\hat{\beta}^2S_{xx}=S_{yy}-2\frac{S_{xy}}{S_{xx}}S_{xy}+\left(\frac{S_{xy}}{S_{xx}}\right)^2S_{xx}=S_{yy}-\frac{{S_{xy}}^2}{S_{xx}}\end{aligned}$$

A4.2　誤差の分散の推定

回帰係数の推定値（A4.5）式を残差平方和 S_eの式に代入すると，上のように残差平方和の最小値 $S_{e,\min}$の式が得られる．誤差 ε_kの不偏分散 σ^2は，$S_{e,min}$をその自由度で割れば計算できる**【⇒第 16 講 e 節】**．残差の個数は n だが，回帰式の制約があるため，自由に値をとれる残差の個数は$(n-2)$個なので**【⇒第 16 講 e 節 Note】**，不偏分散は以下のようになる（自由度は$(n-2)$）．

$$\hat{\sigma}^2=\frac{\sum{\varepsilon_k}^2}{n-2}=\frac{S_{e,\min}}{n-2} \qquad \text{(A4.6)}$$

A4.3　回帰係数の区間推定

A4.1 節で，正規方程式を解くことにより，回帰係数 βの点推定値 $\hat{\beta}$を求められることを示したが，回帰係数βの信頼区間は，誤差が最小二乗法の適用条件を満たしているとき，$\hat{\beta}$が正規分布をすることを使って求めることができるので，概要を示しておく．

誤差 ε_kがX_kによらず同じ分散 σ^2の正規分布をしている，

$$\varepsilon_k \sim N(0,\sigma^2)$$

という仮定と，回帰モデルが線形である，

$$Y_k = \alpha + \beta X_k + \varepsilon_k$$

という仮定から，回帰係数 α，β の推定値 $\hat{\alpha},\hat{\beta}$ は正規分布をすることが示せる．また，$\hat{\alpha}$, $\hat{\beta}$の分散と共分散は以下のようになる．

$$V(\hat{\alpha}) = \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right)\sigma^2, V(\hat{\beta}) = \frac{\sigma^2}{S_{xx}}, Cov(\hat{\alpha},\hat{\beta}) = -\frac{\bar{X}}{S_{xx}}\sigma^2 \qquad \text{(A4.7)}$$

ところで，$\hat{\beta}$ の期待値はβ なので，以下の関係が成立することがわかる．

$$\hat{\beta} \sim N(\beta, \frac{\sigma^2}{S_{xx}}) \qquad \text{(A4.8)}$$

このとき，分散の値が確定していれば，**基本編（第 4 講）**で説明した分散既知の場合の推定や検定がそのまま適用できるが，残念ながらσ^2という局外母数がある．だが，この状況は第 4 講の「未知の母数が 2 個ある場合」とほぼ同じである．そこで，第 4 講と同様の手順で未知の母数 σ^2をその推定量 $\hat{\sigma}^2 = S_{e,min}/(n-2)$ で置き換えた上，基準化統計量をつくると，以下の関係が成立することを示すことができる．

$$t = \frac{\hat{\beta} - \beta}{\frac{\hat{\sigma}}{\sqrt{S_{xx}}}} \sim \text{自由度 } (n-2)\text{の}t\text{分布}$$

あとは，基本編で信頼区間を導いたときと全く同じ手順で，統計量 t の 95%信頼区間を求め（以下の Pr{　} の中の不等式），

$$\Pr\{-t_{n-2}(0.025) \le t \le t_{n-2}(0.025)\} = 0.95$$

Pr{　} の中の不等式を β について解くことにより，β の 95%信頼区間が得られる．

$$\hat{\beta} - t_{n-2}(0.025)\frac{\hat{\sigma}}{\sqrt{S_{xx}}} \le \beta \le \hat{\beta} + t_{n-2}(0.025)\frac{\hat{\sigma}}{\sqrt{S_{xx}}} \qquad (A4.9)$$

なお，$t_{n-2}(0.025)$ は自由度$(n-2)$の t 分布の上位 2.5%点である．

A4.4　回帰式の値の区間推定

任意のX_0に対する回帰直線上の点の Y座標 η_0,

$$\eta_0 = \alpha + \beta X_0$$

に対する区間推定は，η_0 の点推定値

$$\hat{Y}_0 = \hat{\alpha} + \hat{\beta} X_0$$

の期待値と分散を評価することで求めることができる．ここで,

$$V(\hat{\alpha} + \hat{\beta} X_0) = V(\hat{\alpha}) + 2X_0 Cov(\hat{\alpha}, \hat{\beta}) + {X_0}^2 V(\hat{\beta})$$

なので，(A4.7)から

$$= \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right)\sigma^2 - 2X_0 \frac{\bar{X}}{S_{xx}}\sigma^2 + {X_0}^2 \frac{\sigma^2}{S_{xx}} = \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}}\right)\sigma^2$$

となる．したがって,

$$\hat{Y}_0 = \hat{\alpha} + \hat{\beta} X_0 \sim N(\alpha + \beta X_0, \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}}\right)\sigma^2)$$

が成立する．そこで，A4.3 節のβの区間推定の問題と全く同じ手順で，σ^2をその推定値$\hat{\sigma}^2 = S_{e,min}/(n-2)$で置き換えた上，基準化統計量をつくり，それが自由度$(n-2)$のt分布をすることを利用して，以下のように信頼区間を得ることができる．

$$(\hat{\alpha} + \hat{\beta} X_0) - t_{n-2}(0.025)\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}}}\hat{\sigma} \leq \eta_0$$

$$\leq (\hat{\alpha} + \hat{\beta} X_0) + t_{n-2}(0.025)\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}}}\hat{\sigma} \qquad (A4.10)$$

信頼区間の下限，上限を見ると$(X_0 - \bar{X})^2$という項がある．そのため，信頼区間は X が $\bar{X}$ から離れるほど幅が広くなり，下限点，上限点を結ぶと双曲線になる．

A4.5　Yの予測区間

任意のX_0に対する回帰直線上の点でなく，任意の X_0 に対する Y，つまり実際の観測値がどんな値をとるかを考えてみよう．この場合 Y は，A4.4 節で求めた回帰直線上の点 Y_0にさらに誤差 ε が加わって以下の形になる．

$$Y = \eta_0 + \varepsilon = \alpha + \beta X_0 + \varepsilon$$

Y の点推定値は，回帰直線上の点 η_0 と同じく

$$\hat{Y}_0 = \hat{\alpha} + \hat{\beta} X_0$$

なので $\hat{Y}_0$ の期待値は $\alpha + \beta X_0$だが，分散は誤差 ε がある分，大きくなり

$$\left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}}\right)\sigma^2$$

となる．あとは A4.4 とまったく同じ手順で，以下の信頼区間を得ることができる．

$$(\hat{\alpha} + \hat{\beta} X_0) - t_{n-2}(0.025)\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}}}\hat{\sigma} \le Y$$

$$\le (\hat{\alpha} + \hat{\beta} X_0) + t_{n-2}(0.025)\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}}}\hat{\sigma} \qquad (A4.11)$$

回帰直線上の点の信頼区間に比べると少しだけ幅が広くなっている．これを**予測区間**と呼んでいる．

A4.6　てこ比

Xの特定の値 X_iに対する回帰直線上の点 $(X_i, \hat{Y}_i)$ の Y座標 $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ は(14.5)式と(13.6)式を使って変形すると，Y_1，Y_2，Y_3，…を使って以下のように表せる．

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i = \bar{Y} + \hat{\beta}(X_i - \bar{X}) = \bar{Y} + \frac{S_{xy}}{S_{xx}}(X_i - \bar{X})$$

$$= \frac{1}{n}\sum_{j=1}^{n} Y_j + \frac{(X_i - \bar{X})}{S_{xx}}\sum_{j=1}^{n}(X_j - \bar{X})(Y_j - \bar{Y}) = \frac{1}{n}\sum_{j=1}^{n} Y_j + \frac{(X_i - \bar{X})}{S_{xx}}\sum_{j=1}^{n}(X_j - \bar{X})Y_j$$

$$= \sum_{j=1}^{n}\left(\frac{1}{n} + \frac{(X_i - \bar{X})(X_j - \bar{X})}{S_{xx}}\right)Y_j = \sum_{j=1}^{n} h_{ij} Y_j$$

つまり，

$$\hat{Y}_i = h_{i1}Y_1 + h_{i2}Y_2 + \cdots + h_{ii}Y_i + \cdots + h_{in}Y_n,$$

$$h_{ij} = \frac{1}{n} + \frac{(X_i - \bar{X})(X_j - \bar{X})}{S_{xx}}, \text{ 特に } h_{ii} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{S_{xx}}$$

ここで，仮に i 番目の測定値の Yの値が 1 増えて Y_iから (Y_i+1)になったとしよう．このとき，$\hat{Y}_i$ は h_{ii}だけ大きくなる．h_i のことをてこ比と呼ぶが，てこ比 h_{ii}が大きい Y_iは推定に対する影響が大きく，わずかな値の違いで回帰式が大きく変わる可能性があるため，注意が必要である．h_{ii}は$(X_i-\bar{X})^2$にほぼ比例することからわかるように，外側の測定値ほど回帰直線の推定に及ぼす影響が大きく，外れ値にはことのほか注意が必要なことがわかる．

A4.7　$S_{yy}=S_R+S_e$の証明

はじめに 3 つの変動の定義を再掲する（加算の範囲は$k=1,\ \cdots,\ n$；図 14.6）．

① Yの変動（全変動）　　$S_{yy}=\sum(Y_k-\bar{Y})^2$

② 回帰式の変動　　$S_R=\sum(\hat{Y}_k-\bar{Y})^2$

③ 回帰式からの変動　　$S_e=\sum {e_k}^2=\sum(Y_k-\hat{Y}_k)^2=\sum(Y_k-(\hat{\alpha}+\hat{\beta}X_k))^2$

Y_kは測定値なので $Y_k=\hat{\alpha}+\hat{\beta}X_k+e_k$と表せる．また，$\hat{Y}_k$ は X_kにおける回帰式の y 座標なので $\hat{Y}_k=\hat{\alpha}+\hat{\beta}X_k$である．

ところで，一般に以下の関係が成立していると，二乗を「二乗の和」に分解できる．

$$a^2=(b+c)^2=b^2+2bc+c^2=b^2+c^2\quad(bc=0\ \text{のとき})$$

つまり，$bc=0$ となるよう a を$a=b+c$の形に分解できると計算がシンプルになる．そこで，$a=Y_k-\bar{Y}$ と考えて，$bc=0$ となるよう$b=(Y_k-\bar{Y})-e_k=\left(\hat{\alpha}+\hat{\beta}X_k-\bar{Y}\right)$，$c=e_k$として，$S_{yy}$ を展開してみる．

$$S_{yy}=\sum(Y_k-\bar{Y})^2=\sum(b+c)^2=\sum\left(\left(\hat{\alpha}+\hat{\beta}X_k-\bar{Y}\right)+e_k\right)^2$$

$$=\sum\{((\hat{\alpha}+\hat{\beta}X_k)-\bar{Y})^2+2e_k((\hat{\alpha}+\hat{\beta}X_k)-\bar{Y})+{e_k}^2\}$$

ところで，(A4.2)式から$\sum e_k=0$，(A4.3)式から$\sum e_kX_k=0$なので，第 2 項は

$$\sum e_k((\hat{\alpha}+\hat{\beta}X_k)-\bar{Y})=(\hat{\alpha}-\bar{Y})\sum e_k+\hat{\beta}\sum e_kX_k=(\hat{\alpha}-\bar{Y})\cdot 0+\hat{\beta}\cdot 0=0$$

である．よって，

$$S_{yy}=\sum\left((\hat{\alpha}+\hat{\beta}X_k)-\bar{Y}\right)^2+\sum {e_k}^2$$

$$=\sum(\hat{Y}_k-\bar{Y})^2+\sum {e_k}^2=S_R+S_e\qquad\text{(A4.12)}$$

つまり，以下の関係が成立することがわかる．

全変動＝（回帰式の変動）＋（回帰モデルからの変動）

A4.8　相関係数と寄与率の関係

A4.1 節で求めた$S_e = S_{yy} - \frac{{S_{xy}}^2}{S_{xx}}$という関係（$S_e = S_{e,min}$のとき）と，A4.7 節で証明した平方和間の関係（$S_{yy} = S_R + S_e$）を使うと，寄与率は

$$\frac{\text{回帰式の変動}}{\text{全変動}} = \frac{S_R}{S_{yy}} = \frac{S_{yy} - S_e}{S_{yy}} = 1 - \frac{S_e}{S_{yy}} = 1 - \frac{1}{S_{yy}}\left(S_{yy} - \frac{{S_{xy}}^2}{S_{xx}}\right) = \frac{{S_{xy}}^2}{S_{xx}S_{yy}}$$

となる．これは第 13 講の r^2の式（(13.8)式）と等しい．したがって，以下の重要な関係が成立することがわかる．

$$\boldsymbol{r^2 = \text{寄与率} = \frac{\text{回帰式の変動}}{\text{全変動}}} = \frac{S_R}{S_{yy}} \qquad \text{(A4.13)}$$

A4.9　帰無仮説「β =0」に対する検定と分散分析の関係

単回帰の場合は，回帰式の残差平方和S_Rの自由度は，

$$\phi_R = n_p - 1 = 2 - 1 = 1$$

回帰残差 eの平方和の自由度は，

$$\phi_e = n - n_p = n - 2$$

なので，「回帰式に意味がない」という帰無仮説「${\sigma_R}^2 = \sigma^2$」に対する検定統計量 Fは，

$$F = \frac{S_R/\phi_R}{S_e/\phi_e} = \frac{S_R}{S_e/(n-2)}$$

となり，自由度1，$(n-2)$のF分布に従う．

一方，(14.8)式から，

$$z = \frac{\hat{\beta} - \beta}{\sqrt{\hat{\sigma}^2/S_{xx}}} \sim t_{n-2}$$

なので，帰無仮説が「$\beta = 0$」のとき

$$z^2 = \frac{\hat{\beta}^2}{\frac{\hat{\sigma}^2}{S_{xx}}} = \frac{(\frac{S_{xy}}{S_{xx}})^2}{\frac{S_e/(n-2)}{S_{xx}}} = \frac{\frac{{S_{xy}}^2}{S_{xx}}}{S_e/(n-2)} = \frac{S_R}{S_e/(n-2)} = F$$

が成立する．したがって，単回帰の場合，回帰による変動と誤差が等しいという帰無仮説「${\sigma_R}^2 = \sigma^2$」に対する F検定と，帰無仮説「$\beta = 0$」に対する t 分布を使った検定は等価になる．

付録 5　ROC 曲線下面積の意味の証明

ROC 曲線下面積の意味は簡単な積分の計算で証明することができる．発症群を[d]，非発症群を[h]，両群からランダムに選んだ値を $x_{[d]}$，$x_{[h]}$とすると，前者のほうが大きい確率 c は以下のように表せる．

$$c = \int_{-\infty}^{\infty} P(x_{[h]} = z) \cdot P(x_{[d]} > z)\, dz$$

両群の確率密度関数を f_h，f_d とすると，T_p, F_p の定義から

$$Tp = \int_{z}^{\infty} f_d(x)dx = T_p(z)\,,\quad Fp = \int_{z}^{\infty} f_h(x)dx = F_p(z)$$

と表せる．高校で習った置換積分を使って（一番最後の変形で）

$$c = \int_{-\infty}^{\infty} f_h(z) \cdot \{\int_{z}^{\infty} f_d(x)\,dx\}\, dz = \int_{-\infty}^{\infty} f_h(z) \cdot T_p(z)\, dz$$

$$= \int_{-\infty}^{\infty} T_p(z) \cdot f_h(z)\, dz = \int_{-\infty}^{\infty} T_p(z) \cdot \frac{dF_p}{dz} dz = \int_{0}^{1} T_p(z) dF_p(z)$$

と変形できる．最後の式は T_p を F_p（ROC 曲線を描いたときの横軸の変数）で積分することなので，ROC 曲線の下の面積にほかならない．

【参考図書】

Green DM and Swets JA (1989): Signal Detection Theory and Psycophysics. Peninsula Publishing.

Pepe MS (2004): The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press.

付表　χ^2分布の%点

	下位%点			上位%点		
自由度	0.01	0.025	0.05	0.05	0.025	0.01
1	0.00016	0.00098	0.0039	3.841	5.024	6.635
2	0.0201	0.0506	0.103	5.991	7.378	9.210
3	0.115	0.216	0.352	7.815	9.348	11.345
4	0.297	0.484	0.711	9.488	11.143	13.277
5	0.554	0.831	1.145	11.070	12.833	15.086
6	0.872	1.237	1.635	12.592	14.449	16.812
7	1.239	1.690	2.167	14.067	16.013	18.475
8	1.646	2.180	2.733	15.507	17.535	20.090
9	2.088	2.700	3.325	16.919	19.023	21.666
10	2.558	3.247	3.940	18.307	20.483	23.209
11	3.053	3.816	4.575	19.675	21.920	24.725
12	3.571	4.404	5.226	21.026	23.337	26.217
13	4.107	5.009	5.892	22.362	24.736	27.688
14	4.660	5.629	6.571	23.685	26.119	29.141
15	5.229	6.262	7.261	24.996	27.488	30.578
16	5.812	6.908	7.962	26.296	28.845	32.000
17	6.408	7.564	8.672	27.587	30.191	33.409
18	7.015	8.231	9.390	28.869	31.526	34.805
19	7.633	8.907	10.117	30.144	32.852	36.191
20	8.260	9.591	10.851	31.410	34.170	37.566
21	8.897	10.283	11.591	32.671	35.479	38.932
22	9.542	10.982	12.338	33.924	36.781	40.289
23	10.196	11.689	13.091	35.172	38.076	41.638
24	10.856	12.401	13.848	36.415	39.364	42.980
25	11.524	13.120	14.611	37.652	40.646	44.314
26	12.198	13.844	15.379	38.885	41.923	45.642
27	12.879	14.573	16.151	40.113	43.195	46.963
28	13.565	15.308	16.928	41.337	44.461	48.278
29	14.256	16.047	17.708	42.557	45.722	49.588
30	14.953	16.791	18.493	43.773	46.979	50.892
40	22.164	24.433	26.509	55.758	59.342	63.691
50	29.707	32.357	34.764	67.505	71.420	76.154
100	70.065	74.222	77.929	124.342	129.561	135.807