

統計学的推測（入門編）

汪 金芳

横浜市立大学・データサイエンス学部

July 20, 2018

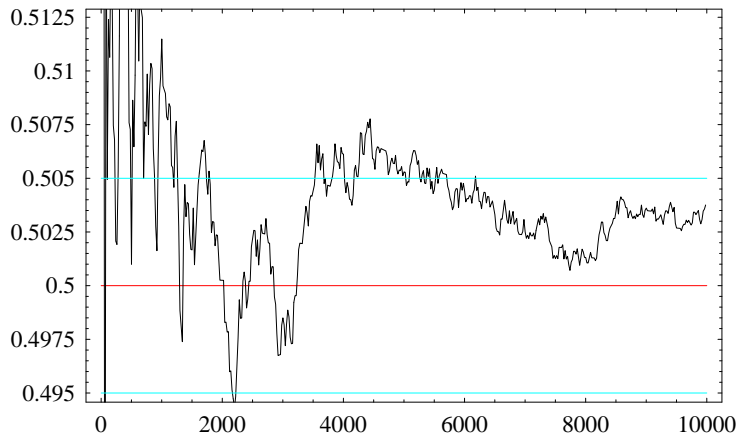
「統計学的推測（入門編）」目次

- 1 第1講：大数の弱法則
- 2 第2講：中心極限定理
- 3 第3講：正規分布からの標本抽出- χ^2 分布
- 4 第4講：正規分布からの標本抽出: t 分布
- 5 第5講：区間推定: 正規分布の平均の場合
- 6 第6講：区間推定: 正規分布の分散の場合
- 7 第7講：仮説検定
- 8 第8講：仮説検定: 尤度比検定
- 9 第9講：仮説検定: 信頼区間による方法
- 10 第10講：標本数の決定
- 11 第11講：生存時間解析入門

第1講：大数の弱法則

数値実験：ベルヌーイ試行

正しいコインを、独立に $n = 1, 2, \dots, 10000$ 回投げた場合



横軸：投げる回数；縦軸：表の出る相対頻度

大数の弱法則 (続き)

定理 (大数の弱法則; WEAK LAW OF LARGE NUMBERS)

① 独立性: 確率変数 X_1, X_2, \dots, X_n が互いに独立

② 平均の同一性:

$$\mu = E(X_i), i = 1, 2, \dots, n$$

③ 分散の有限性:

$$\sigma_i^2 = V(X_i) \leq \sigma^2, i = 1, 2, \dots, n$$

このとき、任意の $\epsilon > 0$ に対して

$$\lim_{n \rightarrow \infty} P \left(\left| \frac{X_1 + X_2 + \dots + X_n}{n} - \mu \right| \geq \epsilon \right) = 0 \quad (1)$$

(1) が成り立つとき、 \bar{X} が μ に 確率収束 *converge in probability* するという。

大数の弱法則: 証明

$$Y = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

とおくと、 $E(Y) = \mu$ となる。また独立性より

$$V(Y) = \frac{\sigma_1^2 + \cdots + \sigma_n^2}{n^2} \leq \frac{\sigma^2 + \cdots + \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

を得る。チェビシエフの不等式

$$P(|Y - E(Y)| \geq \epsilon) \leq \frac{V(Y)}{\epsilon^2}$$

より

$$P(|Y - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$$

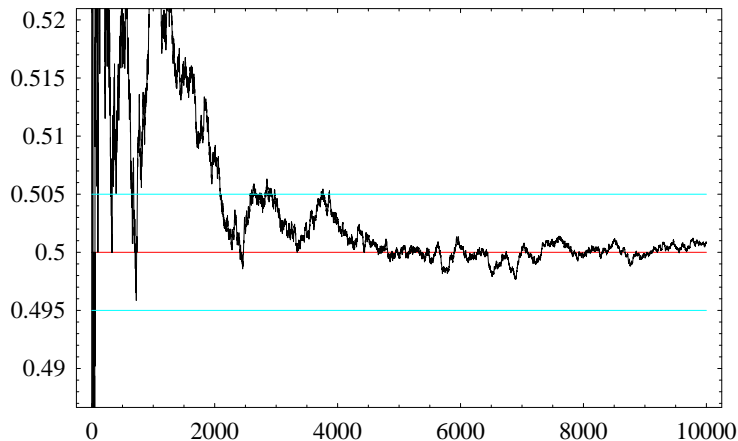
を得る。したがって、

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right) = 0$$

数の強法則: コインを投げる場合

次々に正しいコインを $n = 10000$ 回投げた場合

横軸：投げる回数；縦軸：表の出る相対頻度



チェビシェフの不等式

定理 (チェビシェフの不等式)

X の平均を $\mu = E(X)$, 偶数次の中心モーメントを

$$\nu_{2k} = E(X - \mu)^{2k}$$

とする. このとき, 任意の $\epsilon > 0$ に対して, 次が成り立つ

$$P(|X - \mu| \geq \epsilon) \leq \frac{\nu_{2k}}{\epsilon^{2k}} \quad (2)$$

証明: $D = \{x | x - \mu| \geq \epsilon\}$ とする。

$$\begin{aligned} \nu_{2k} &= \int_{-\infty}^{\infty} (x - \mu)^{2k} f(x) dx \geq \int_D (x - \mu)^{2k} f(x) dx \\ &\geq \int_D \epsilon^{2k} f(x) dx \geq \epsilon^{2k} \int_D f(x) dx \\ &\geq \epsilon^{2k} P(|X - \mu| \geq \epsilon) \end{aligned}$$

標本平均の4次のモーメント

定理 (標本平均の4次のモーメント)

- X_1, X_2, \dots, X_n が互いに独立で,
- $E(X_i) = \mu, V(X_i) = \sigma^2, E(X_i - \mu)^4 = \nu_4, (i = 1, \dots, n)$ とすると, 次が成り立つ。

$$E(\bar{X} - \mu)^4 = \frac{1}{n^2} \left[\frac{1}{n} \nu_4 + 3 \left(1 - \frac{1}{n}\right) \sigma^4 \right] \quad (3)$$

証明:

$$\begin{aligned} (\bar{X} - \mu)^4 &= \left[\frac{1}{n} \sum_{i=1}^n (X_i - \mu) \right]^4 \\ &= \frac{1}{n^4} \left\{ \sum_{i=1}^n (X_i - \mu)^4 + \sum_{i \neq j} {}_4C_2 (X_i - \mu)^2 (X_j - \mu)^2 \right. \\ &\quad + \sum_{i \neq j \neq k \neq \ell} {}_4C_1 (X_i - \mu) (X_j - \mu) (X_k - \mu) (X_\ell - \mu) \\ &\quad \left. + \sum_{i \neq j} {}_4C_1 (X_i - \mu)^3 (X_j - \mu) \right\} \end{aligned}$$

両辺期待値を取ると, $E(\bar{X} - \mu)^4 = \frac{1}{n^4} (n\nu_4 + n {}_4C_2 {}_4C_2 \sigma^4)$

定理 (大数の強法則; STRONG LAW OF LARGE NUMBERS)

- 1 確率変数 X_1, X_2, \dots, X_n : 互いに独立で同一分布に従い,
- 2 さらに,

$$\begin{cases} \mu = E(X_i) \\ \sigma^2 = V(X_i) \\ \nu_4 = E(X_i - \mu)^4 \end{cases} \quad i = 1, 2, \dots, n$$

と仮定する。このとき、次が成り立つ。

$$P\left(\lim_{n \rightarrow \infty} \frac{X_1 + X_2 + \dots + X_n}{n} = \mu\right) = 1 \quad (4)$$

(4) が成り立つとき、 \bar{X} が μ に 概収束 *converge almost surely* (*converge with probability 1*) という。

大数の強法則の証明

- 評価すべき事象は

$$\lim_{n \rightarrow \infty} \bar{X} = \mu$$

である。

- この事象は次の事象と同等である。任意の $\epsilon > 0$ に対して、自然数 N_ϵ が存在し、全ての $n > N_\epsilon$ に対して

$$|\bar{X} - \mu| < \epsilon, \quad \forall \epsilon > 0, n > N_\epsilon$$

という事象である。

- この事象の余事象は、ある ϵ が存在し、この ϵ に対して、どんな大きい N をとっても、 $n > N$ が存在し、次が満たされる事象

$$|\bar{X} - \mu| \geq \epsilon, \quad \text{for some } n > N$$

である。

大数の強法則の証明 (つづき)

一方、 $n > N > \nu_4$ のとき、チェビシエフの不等式より

$$\begin{aligned} P_n &= P(|\bar{X} - \mu| \geq \epsilon) \\ &\leq \frac{E(\bar{X} - \mu)^4}{\epsilon^4} \\ &= \frac{1}{n^2 \epsilon^4} \left[\frac{1}{n} \nu_4 + 3 \left(1 - \frac{1}{n}\right) \sigma^4 \right] \\ &< \frac{1 + 3\sigma^4}{\epsilon^4} \frac{1}{n^2} \end{aligned}$$

したがって、1つの $\epsilon > 0$ と任意の N に対して、

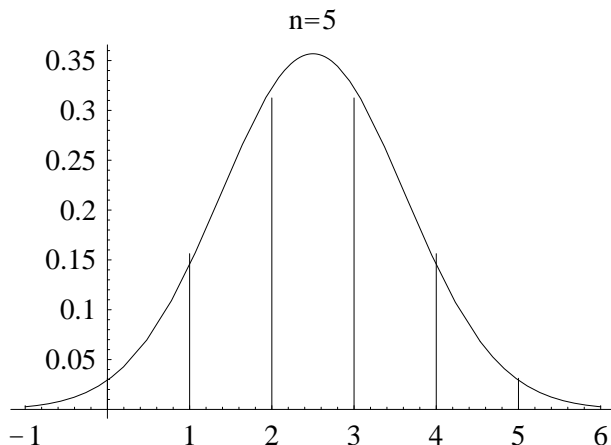
$$\begin{aligned} P\left(\lim_{n \rightarrow \infty} \bar{X} = \mu\right) &= 1 - P(|\bar{X} - \mu| \geq \epsilon \text{ for some } n > N) \\ &\geq 1 - (P_N + P_{N+1} + \cdots) \\ &\geq 1 - \frac{1 + 3\sigma^4}{\epsilon^4} \left\{ \frac{1}{N^2} + \frac{1}{(N+1)^2} + \cdots + \right\} \\ &\xrightarrow{N \rightarrow \infty} 1 \end{aligned}$$

最後の極限は、無限級数 $\sum_{i=1}^{\infty} 1/N^2$ の収束性による。

第 2 講：中心極限定理

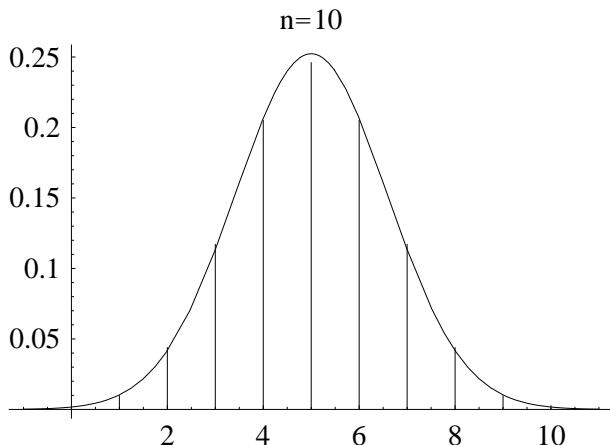
二項分布の場合

X_1, \dots, X_n : 独立で、 $p = 1/2$ の Bernoulli 分布に従い、 $\sum_{i=1}^n X_i$ の分布



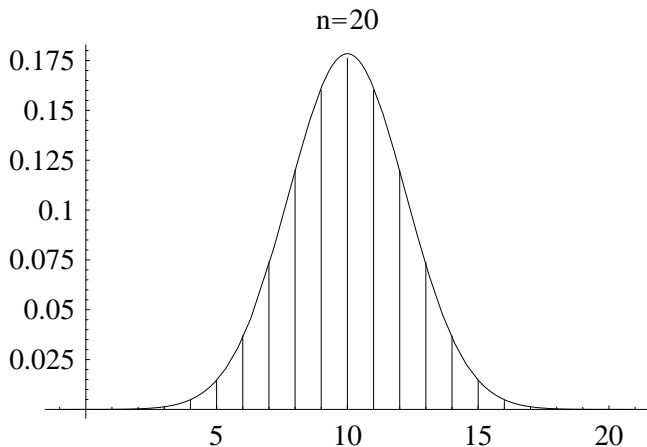
中心極限定理: 二項分布の場合 (つづき)

X_1, \dots, X_n : 独立で、 $p = 1/2$ の Bernoulli 分布に従い、 $\sum_{i=1}^n X_i$ の分布



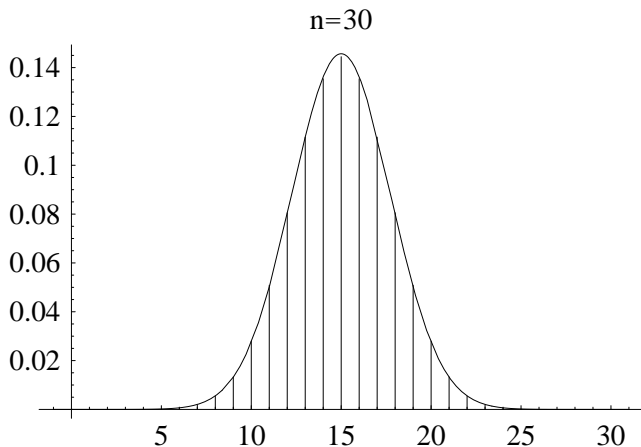
中心極限定理: 二項分布の場合 (つづき)

X_1, \dots, X_n : 独立で、 $p = 1/2$ の Bernoulli 分布に従い、 $\sum_{i=1}^n X_i$ の分布



中心極限定理: 二項分布の場合 (つづき)

X_1, \dots, X_n : 独立で、 $p = 1/2$ の Bernoulli 分布に従い、 $\sum_{i=1}^n X_i$ の分布



定理 (中心極限定理; CENTRAL LIMIT THEOREM)

次の条件の下で

- 1 X_1, \dots, X_n が独立で、同じ分布に従う。
- 2 平均と分散が存在し,

$$E(X_i) = \mu, V(X_i) = \sigma^2, i = 1, \dots, n$$

とする。

このとき、確率変数 $Y = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$ は標準正規分布に弱収束する。すなわち

$$\lim_{n \rightarrow \infty} P \left[\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \leq y \right] = \int_{-\infty}^y \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \quad (5)$$

中心極限定理: 証明 (1/2)

Y の積率母関数が標準正規分布のそれに近づくことを証明する。
 X_i の積率母関数が存在するという、定理より強い条件を仮定する。
 $Y_i = \frac{X_i - \mu}{\sigma}$ とすると、

$$E(Y_i) = 0, V(Y_i) = 1, Y = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i$$

Y_i の積率母関数を $M(t)$ とすると

$$\begin{aligned} M(0) &= E(e^{0Y_i}) = 1 \\ M'(0) &= E(Y_i) = 0 \\ M''(0) &= E(Y_i^2) = V(Y_i) + (EY_i)^2 = 1 \end{aligned}$$

また $\psi(t) = \log M(t)$ とすると、

$$\psi'(t) = \frac{M'(t)}{M(t)}, \quad \psi''(t) = \frac{M''(t)M(t) - [M'(t)]^2}{[M(t)]^2}$$

したがって、 $\psi(0) = 0$, $\psi'(0) = 0$, $\psi''(0) = 1$ となる。

中心極限定理: 証明 (2/2)

一方, ある $|\theta| < |t|$ が存在し、原点の周りで $\psi(t)$ を展開すると

$$\psi(t) = \psi(0) + \frac{\psi'(0)}{1!}t + \frac{\psi''(0)}{2!}t^2 + \frac{\psi'''(\theta)}{3!}t^3 = \frac{1}{2}t^2 + \frac{\psi'''(\theta)}{6}t^3$$

従って、 Y の積率母関数は

$$\begin{aligned} E(e^{Yt}) &= E\left(e^{\frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i t}\right) \\ &= \prod_{i=1}^n E\left(e^{\frac{t}{\sqrt{n}} Y_i}\right) \\ &= [M(t/\sqrt{n})]^n \\ &= \exp\{n \log M(t/\sqrt{n})\} \\ &= \exp\left\{n \left[\frac{1}{2} (t/\sqrt{n})^2 + \frac{\psi'''(\theta)}{6} (t/\sqrt{n})^3\right]\right\} \\ &= \exp\left\{\frac{1}{2}t^2 + \frac{\psi'''(\theta)}{6} \frac{t^3}{\sqrt{n}}\right\} \\ &\xrightarrow{n \rightarrow \infty} \exp\left\{\frac{1}{2}t^2\right\} \end{aligned}$$

二項分布の正規近似

定理 (DE MOIVRE-LAPLACE LIMIT THEOREM)

X が二項分布

$$X \sim Bi(n, p) \quad (6)$$

に従うとき, 任意の実数 $a, b (a \leq b)$ に対して,

$$\lim_{n \rightarrow \infty} P \left[a \leq \frac{X - np}{\sqrt{np(1-p)}} \leq b \right] = \Phi(b) - \Phi(a) \quad (7)$$

が成り立つ。ただし、 $\Phi(x)$ は標準正規分布の分布関数を表す。すなわち、

$$\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy$$

二項分布の正規近似: 証明

X_1, \dots, X_n を独立で、成功する確率が p の Bernoulli 分布に従うならば、

$$\begin{aligned} X &= \sum_{i=1}^n X_i \sim \text{Bi}(n, p) \\ E(X_i) &= p \\ V(X_i) &= p(1-p) \end{aligned}$$

となる。

中心極限定理によって

$$\begin{aligned} \frac{X - np}{\sqrt{np(1-p)}} &= \frac{\sum_{i=1}^n X_i - np}{\sqrt{np(1-p)}} \\ &= \frac{\sqrt{n}(\bar{X} - p)}{\sqrt{p(1-p)}} \\ &\rightarrow N(0, 1) \end{aligned}$$

第3講：正規分布からの標本抽出- χ^2 分布

χ^2 分布

正規分布からの標本抽出を考える。次を仮定する。

$$(A1) \quad X_1, \dots, X_n \sim N(0, 1)$$

(A2) X_1, \dots, X_n : 独立

定理

(A1), (A2) の下で, $Y = X_1^2 + \dots + X_n^2$ の密度関数 $f(y)$ は次の式で与えられる。

$$f(y) = \frac{1}{\Gamma(n/2)} \left(\frac{1}{2}\right)^{\frac{n}{2}} y^{\frac{n}{2}-1} e^{-\frac{y}{2}} I_{(0, \infty)}(y) \quad (8)$$

ただし, $\Gamma(z)$ は次で定義されるガンマ関数である。

$$\Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t} dt \quad (z > 0)$$

Definition

(8) を自由度 n の χ^2 分布の密度関数という。

χ^2 分布: 証明 (1/2)

X_1, \dots, X_n の独立性より

$$\begin{aligned}M_Y(t) &= E[\exp(Yt)] \\&= E\left(\exp\left\{\sum_{i=1}^n X_i^2 t\right\}\right) \\&= \left(E e^{X_1^2 t}\right)^n\end{aligned}$$

一方

$$\begin{aligned}E e^{X_1^2 t} &= \int_{-\infty}^{\infty} e^{x^2 t} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \\&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}(1-2t)} dx \\&= \frac{1}{\sqrt{1-2t}} \quad (1-2t > 0)\end{aligned}$$

従って、

$$M_Y(t) = (1-2t)^{-\frac{n}{2}}$$

χ^2 分布: 証明 (2/2)

直接的に計算すると

$$\begin{aligned}\int_0^\infty e^{yt} f(y) dy &= \int_0^\infty e^{yt} \frac{1}{\Gamma(n/2)} \left(\frac{1}{2}\right)^{\frac{n}{2}} y^{\frac{n}{2}-1} e^{-\frac{y}{2}} dy \\ &= \int_0^\infty \frac{1}{\Gamma(n/2)} \left(\frac{1}{2}\right)^{\frac{n}{2}} y^{\frac{n}{2}-1} e^{-\frac{y}{2}+yt} dy \\ \left(x = \frac{y}{2} - yt\right) &= \int_0^\infty \frac{1}{\Gamma(n/2)} \left(\frac{1}{2}\right)^{\frac{n}{2}} \left(\frac{2}{1-2t}x\right)^{\frac{n}{2}-1} e^{-x} \left(\frac{2}{1-2t}dx\right) \\ &= \left(\frac{1}{1-2t}\right)^{\frac{n}{2}} \frac{1}{\Gamma(n/2)} \int_0^\infty x^{\frac{n}{2}-1} e^{-x} dx \\ &= \left(\frac{1}{1-2t}\right)^{\frac{n}{2}}\end{aligned}$$

これが $M_Y(t)$ に一致する。

- 積率母関数

$$M(t) = (1 - 2t)^{-\frac{n}{2}}$$

より

$$\begin{aligned}M'(t) &= n(1 - 2t)^{-\frac{n}{2}-1} \\M''(t) &= n(n + 2)(1 - 2t)^{-\frac{n}{2}-2}\end{aligned}$$

- 平均 :

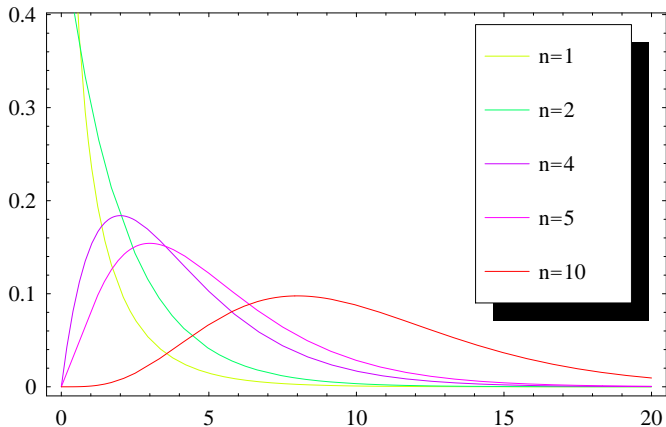
$$E(Y) = M'(0) = n$$

- 分散 :

$$V(Y) = E(X^2) - (EX)^2 = M''(0) - n^2 = 2n$$

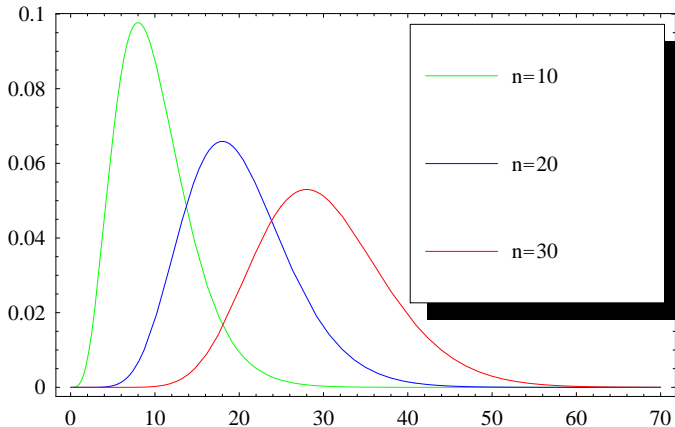
χ^2 分布の様子

χ^2 分布の密度関数



χ^2 分布: 自由度大きい場合

自由度が大きいときの χ^2 分布の密度関数



χ^2 分布の正規近似

- $Z_1 = X_1^2, \dots, Z_n = X_n^2$ が独立に自由度 1 の χ^2 分布に従う
- $E(Z_i) = 1, V(Z_i) = 2$
- $\bar{Z} = \frac{1}{n} \sum_{i=1}^n X_i^2 = Y/n$ とすると
- 中心極限定理により、自由度 n が大きいときに、

$$\frac{\sqrt{n}(Y/n - 1)}{\sqrt{2}} = \frac{\sqrt{n}(\bar{Z} - 1)}{\sqrt{2}} \\ \longrightarrow N(0, 1)$$

$$\implies Y \longrightarrow N(n, 2n)$$

F 分布

定理

(A1) X : 自由度 m の χ^2 分布

(A2) Y : 自由度 n の χ^2 分布

(A3) X, Y : 独立

を仮定する。このとき、 $Z = \frac{X/m}{Y/n}$ の密度関数 $f(z)$ が次の式で与えられる

$$f(z) = \frac{1}{B\left(\frac{m}{2}, \frac{n}{2}\right)} \left(\frac{m}{n}\right)^{\frac{m}{2}} z^{\frac{m}{2}-1} \left(1 + \frac{m}{n}z\right)^{-\frac{m+n}{2}} I_{(0, \infty)}(z) \quad (9)$$

ただし、 $B(x, y)$ は次で定義されるベータ関数である。

$$B(x, y) = \int_0^1 t^{x-1} (1-t)^{y-1} dt = \frac{\Gamma(x) \Gamma(y)}{\Gamma(x+y)}$$

定義

(9) を自由度 (m, n) の F 分布の密度関数という。

F 分布: 証明 (1/3)

- X, Y の独立性から、 X, Y の同時密度関数は

$$f(x, y) = \frac{2^{-\frac{m}{2}}}{\Gamma(m/2)} x^{\frac{m}{2}-1} e^{-\frac{x}{2}} I_{x>0}(x) \times \frac{2^{-\frac{n}{2}}}{\Gamma(n/2)} y^{\frac{n}{2}-1} e^{-\frac{y}{2}} I_{y>0}(y)$$

- 1 対 1 の変数変換

$$\begin{cases} z &= \frac{x/m}{y/n} = \frac{nx}{my} \\ w &= y \end{cases} \Leftrightarrow \begin{cases} x &= \frac{m}{n}zw \\ y &= w \end{cases}$$

のヤコビアン Jacobian は

$$\frac{\partial(x, y)}{\partial(z, w)} = \begin{vmatrix} \frac{\partial x}{\partial z} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial z} & \frac{\partial y}{\partial w} \end{vmatrix} = \begin{vmatrix} \frac{m}{n}w & \frac{m}{n}z \\ 0 & 1 \end{vmatrix} = \frac{m}{n}w$$

F 分布: 証明 (2/3)

したがって

$$\begin{aligned} f(z) &= \int_0^{\infty} g(z, w) dw \\ &= \int_0^{\infty} f\left(\frac{m}{n}zw, w\right) \frac{m}{n} w dz dw \\ &= \int_0^{\infty} \frac{2^{-\frac{m+n}{2}}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{m}{n}zw\right)^{\frac{m}{2}-1} e^{-\frac{m}{2n}zw} w^{\frac{n}{2}-1} e^{-\frac{w}{2}} \frac{m}{n} w dw dz \\ &= \frac{\left(\frac{m}{n}\right)^{\frac{m}{2}} z^{\frac{m}{2}-1}}{2^{\frac{m+n}{2}} \Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \int_0^{\infty} w^{\frac{m+n}{2}-1} e^{-\frac{w(mz+n)}{2n}} dw \end{aligned}$$

変数変換 $t = -\frac{w(mz+n)}{2n}$ すると、

$$w = \frac{2nt}{mz+n}, \quad dw = \frac{2n}{mz+n} dt$$

F 分布: 証明 (3/3)

$$\begin{aligned}\int_0^\infty w^{\frac{m+n}{2}-1} e^{-\frac{w(mz+n)}{2n}} dw &= \int_0^\infty \left(\frac{2nt}{mz+n} \right)^{\frac{m+n}{2}-1} e^{-t} \times \frac{2n}{mz+n} dt \\ &= \left(\frac{2n}{mz+n} \right)^{\frac{m+n}{2}} \int_0^\infty t^{\frac{m+n}{2}-1} e^{-t} dt \\ &= \left(\frac{2n}{mz+n} \right)^{\frac{m+n}{2}} \Gamma\left(\frac{m+n}{2}\right)\end{aligned}$$

従って

$$\begin{aligned}f(z) &= m^{\frac{m}{2}} n^{\frac{n}{2}} \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \frac{z^{\frac{m}{2}-1}}{(mz+n)^{\frac{m+n}{2}}} \\ &= \frac{\left(\frac{m}{n}\right)^{\frac{m}{2}}}{B\left(\frac{m}{2}, \frac{n}{2}\right)} z^{\frac{m}{2}-1} \left(1 + \frac{m}{n}z\right)^{-\frac{m+n}{2}}\end{aligned}$$

Q.E.D.

F 分布の平均・分散

定理

Z が自由度 (m, n) の F 分布に従うとき、次が成り立つ。

$$\text{平均: } E(Z) = \frac{n}{n-2} \quad (n > 2)$$

$$\text{分散: } V(Z) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \quad n > 4$$

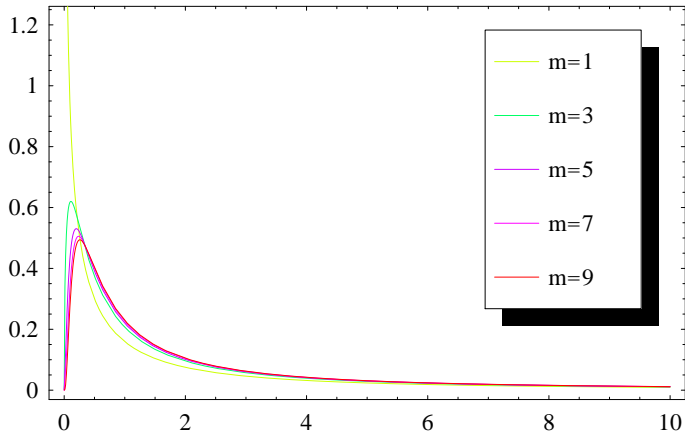
平均の場合の証明: まず

$$\begin{aligned} E\left(\frac{1}{Y}\right) &= \int_0^{\infty} \frac{1}{y} \frac{1}{\Gamma(n/2)} \left(\frac{1}{2}\right)^{n/2} y^{\frac{n}{2}-1} e^{-\frac{y}{2}} dy \\ &= \frac{1}{\Gamma(n/2)} \left(\frac{1}{2}\right)^{n/2} \int_0^{\infty} y^{\frac{n-2}{2}-1} e^{-\frac{y}{2}} dy \\ \left(\frac{y}{2} = w\right) &= \frac{1}{\Gamma(n/2)} \left(\frac{1}{2}\right)^{n/2} \int_0^{\infty} 2^{\frac{n}{2}-2} w^{\frac{n-2}{2}-1} e^{-w} 2 dw \\ &= \frac{1}{2} \frac{\Gamma((n-2)/2)}{\Gamma(n/2)} = \frac{1}{2} \frac{1}{\frac{n}{2}-1} = \frac{1}{n-2} \end{aligned}$$

に注意すると、 $E(Z) = E\left(\frac{X/m}{Y/n}\right) = \frac{n}{m} E(X) E\left(\frac{1}{Y}\right) = \frac{n}{m} m \frac{1}{n-2} = \frac{n}{n-2}$

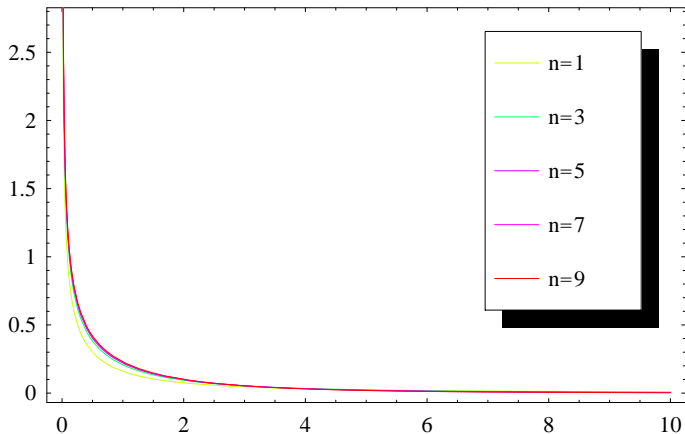
F 分布の密度関数

$Z = \frac{X/m}{Y/n}$, $n = 1$ の場合:



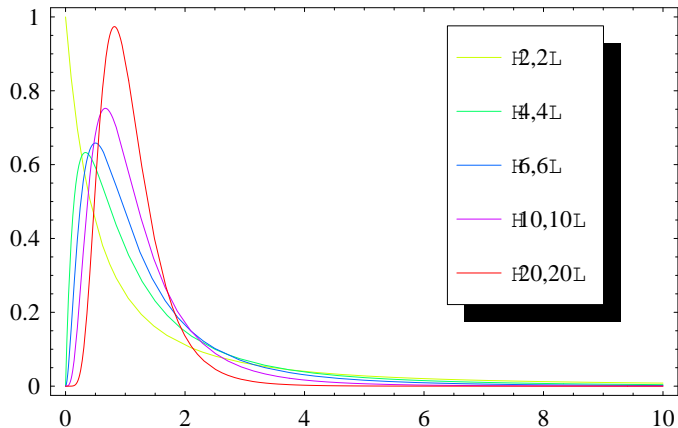
F 分布の密度関数

$Z = \frac{X/m}{Y/n}$, $m = 1$ の場合:



F 分布の密度関数

$$Z = \frac{X/m}{Y/n}, \quad m = n \text{ の場合:}$$



第4講：正規分布からの標本抽出: t 分布

t 分布

t 分布の発見は統計学史上の1つのブレイクスルーである。

定理

(A1) $X \sim N(0, 1)$

(A2) $Y \sim \chi^2(n)$

(A3) X と Y が独立

を仮定する。このとき、 $T = \frac{X}{\sqrt{Y/n}}$ の密度関数 $f(t)$ は次の式で与えられる

$$f(t) = \frac{1}{\sqrt{n}B\left(\frac{n}{2}, \frac{1}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} \quad (10)$$

定義

(10) を自由度 n の t 分布の密度関数という。

t 分布: 証明 (1/3)

- X と Y の独立性から、 X, Y の同時分布

$$f(x, y) = f_X(x) f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \left(\frac{1}{2}\right)^{\frac{n}{2}} \frac{1}{\Gamma(n/2)} y^{\frac{n}{2}-1} e^{-\frac{y}{2}}$$

- 1 \leftrightarrow 1 変数変換

$$\begin{cases} t & = \frac{x}{\sqrt{y/n}} \\ w & = y \end{cases} \iff \begin{cases} x & = \sqrt{\frac{w}{n}} t \\ y & = w \end{cases}$$

ヤコビアン:

$$\begin{vmatrix} \frac{\partial x}{\partial t} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial t} & \frac{\partial y}{\partial w} \end{vmatrix} = \begin{vmatrix} \sqrt{\frac{w}{n}} & \frac{t}{2\sqrt{wn}} \\ 0 & 1 \end{vmatrix} = \sqrt{\frac{w}{n}}$$

t 分布: 証明 (2/3)

したがって

$$\begin{aligned} f(t) &= \int_0^{\infty} g(t, w) dw \\ &= \int_0^{\infty} f\left(\sqrt{\frac{w}{n}}t, w\right) \sqrt{\frac{w}{n}} dw \\ &= \int_0^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\frac{w}{n}t^2} \left(\frac{1}{2}\right)^{\frac{n}{2}} \frac{1}{\Gamma(n/2)} w^{\frac{n}{2}-1} e^{-\frac{w}{2}} \sqrt{\frac{w}{n}} dw \\ &= \frac{1}{\sqrt{2\pi}} \left(\frac{1}{2}\right)^{\frac{n}{2}} \frac{1}{\sqrt{n} \Gamma(n/2)} \int_0^{\infty} w^{\frac{n}{2}-\frac{1}{2}} e^{-\frac{w}{2}-\frac{w}{2n}t^2} dw \end{aligned}$$

次の変数変換を考える

$$s = \frac{w}{2} + \frac{w}{2n}t^2, \quad w = \frac{s}{1/2 + t^2/(2n)}, \quad dw = \left(\frac{1}{2} + \frac{t^2}{2n}\right)^{-1} ds$$

t 分布: 証明 (3/3)

したがって、

$$\begin{aligned} f(t) &= \left(\frac{1}{2}\right)^{\frac{n}{2}} \frac{1}{\sqrt{2\pi}\sqrt{n}\Gamma(n/2)} \left(\frac{1}{2} + \frac{t^2}{2n}\right)^{-\frac{1}{2}-\frac{n}{2}} \int_0^\infty s^{\frac{n}{2}-\frac{1}{2}} e^{-s} ds \\ &= \left(\frac{1}{2}\right)^{\frac{n}{2}} \frac{1}{\sqrt{2\pi}\sqrt{n}\Gamma(n/2)} \left(\frac{1}{2} + \frac{t^2}{2n}\right)^{-\frac{1}{2}-\frac{n}{2}} \Gamma\left(\frac{n}{2} + \frac{1}{2}\right) \end{aligned}$$

次の式

$$B\left(\frac{n}{2}, \frac{1}{2}\right) = \frac{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{1}{2}\right)}{\Gamma\left(\frac{n}{2} + \frac{1}{2}\right)} = \frac{\sqrt{\pi}\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n}{2} + \frac{1}{2}\right)}$$

に注意すると, t 分布の密度関数は

$$\begin{aligned} f(t) &= \left(\frac{1}{2}\right)^{\frac{n}{2}+\frac{1}{2}} \frac{1}{\sqrt{n}B\left(\frac{n}{2}, \frac{1}{2}\right)} \left(\frac{1}{2} + \frac{t^2}{2n}\right)^{-\frac{1}{2}-\frac{n}{2}} \\ &= \frac{1}{\sqrt{n}B\left(\frac{n}{2}, \frac{1}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} \end{aligned}$$

t 分布: 自由度大きいとき

定理

自由度が大きいとき, t 分布は正規分布に近づく.

証明: 次の極限

$$\left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} = \left[\left(1 + \frac{t^2}{n}\right)^{\frac{n}{t^2}}\right]^{\frac{t^2}{n} \times \left(-\frac{n+1}{2}\right)}$$
$$\xrightarrow{n \rightarrow \infty} e^{-\frac{t^2}{2}}$$

に注意すると, 次の成り立つ

$$f(t) = \frac{1}{\sqrt{n} B\left(\frac{n}{2}, \frac{1}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$
$$\xrightarrow{n \rightarrow \infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

別証: 大数の法則による (略)

t 分布: 平均分散

定理

T が自由度 n の t 分布に従うとき, 次が成り立つ。

$$E(T) = 0 \quad (11)$$

$$\text{Proof: } E(T) = E\left(\frac{X}{\sqrt{Y/n}}\right) = E(X) \cdot E\left(\frac{1}{\sqrt{Y/n}}\right) = 0 \cdot E\left(\frac{1}{\sqrt{Y/n}}\right) = 0$$

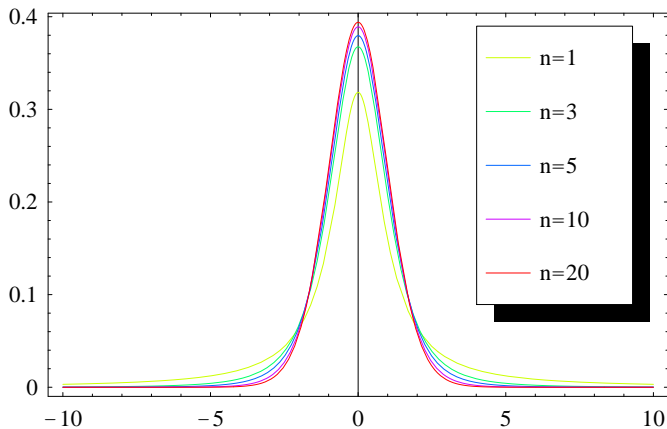
定理

T が自由度 n の t 分布に従うとき, 次が成り立つ。

$$V(T) = \frac{n}{n-2} \quad (12)$$

$$\begin{aligned} \text{Proof: } V(T) &= E(T^2) = nE\left(\frac{X^2}{Y}\right) \\ &= nE(X^2) \cdot E\left(\frac{1}{Y}\right) = n \cdot 1 \cdot \frac{1}{n-2} = \frac{n}{n-2} \end{aligned}$$

t 分布の様子



定理

X_1, X_2, \dots, X_n が独立に $N(0, 1)$ に従うとき、次が成り立つ。

$$\bar{X} = n^{-1} \sum_{i=1}^n X_i \sim N(0, 1/n) \quad (13)$$

$$S^2 = \sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi^2(n-1) \quad (14)$$

$$\bar{X} \text{ is independent of } S^2 \quad (15)$$

証明 まず、(1) を証明する。 \bar{X} の積率母関数を $M(t)$ とする。

$$\begin{aligned} M(t) &= E\left(e^{\bar{X}t}\right) = E\left(e^{\frac{1}{n} \sum_{i=1}^n X_i t}\right) = \prod_{i=1}^n E\left(e^{\frac{t}{n} X_i}\right) \\ &= \left[\exp\left(t^2/2n^2\right)\right]^n = \exp\left(t^2/2n\right) \end{aligned}$$

これは $N(0, 1/n)$ の積率母関数と一致する。

\bar{X} と S^2 の独立性 (1/3)

次に (2), (3) を証明する。次の変数変換を考える

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{n-1} \\ Y_n \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \cdots & -\frac{n-1}{\sqrt{n(n-1)}} \\ \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_{n-1} \\ X_n \end{bmatrix}$$
$$= \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_{n-1} \\ \mathbf{a}_n \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_{n-1} \\ X_n \end{bmatrix} = \mathbf{A} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_{n-1} \\ X_n \end{bmatrix}$$

明らかに,

- $\mathbf{a}_i \mathbf{a}_i^t = 1, \quad i = 1, \dots, n$
- $\mathbf{a}_i \mathbf{a}_j^t = 0, \quad i \neq j = 1, \dots, n$

\bar{X} と S^2 の独立性 (2/3)

すると、次のことが分かる

- Y_1, \dots, Y_n : 正規分布に従う
- $E(Y_i) = 0, V(Y_i) = 1, i = 1, \dots, n$
- $E(Y_i Y_j) = 0, i \neq j = 1, \dots, n$
すなわち、 Y_1, \dots, Y_n は互いに独立である
- $\sum_{i=1}^n X_i^2 = \sum_{i=1}^n Y_i^2$:

$$\begin{aligned}\sum_{i=1}^n Y_i^2 &= [Y_1, Y_2, \dots, Y_n] [Y_1, Y_2, \dots, Y_n]^t \\ &= ([X_1, X_2, \dots, X_n] A^t) (A [X_1, X_2, \dots, X_n]^t) \\ &= [X_1, X_2, \dots, X_n] (A^t A) [X_1, X_2, \dots, X_n]^t \\ &= [X_1, X_2, \dots, X_n] I_{n \times n} [X_1, X_2, \dots, X_n]^t \\ &= \sum_{i=1}^n X_i^2\end{aligned}$$

\bar{X} と S^2 の独立性 (3/3)

従って、 $\sum_{i=1}^{n-1} Y_i^2$ は $\chi^2(n-1)$ に従い、また Y_n と独立。
ところで、

$$Y_n = \sqrt{n}\bar{X}$$

なので、 $\sum_{i=1}^{n-1} Y_i^2$ が次のように書き直すことができる。

$$\begin{aligned}\sum_{i=1}^{n-1} Y_i^2 &= \sum_{i=1}^n X_i^2 - Y_n^2 \\ &= \sum_{i=1}^n X_i^2 - n(\bar{X})^2 \\ &= \sum_{i=1}^n (X_i - \bar{X})^2 \\ &= S^2\end{aligned}$$

Q.E.D.

第5講：区間推定：正規分布の平均の場合

正規分布の平均の信頼区間

母数の区間推定 について考える。

定義 (信頼区間; CONFIDENCE INTERVAL)

(A1) 無作為標本: $X_1, X_2, \dots, X_n \stackrel{i.i.d.}{\sim} f(x|\theta)$

(A2) 密度関数 $f(x|\theta)$ が未知の母数 θ に依存するが, 関数 $f(\cdot|\cdot)$ は既知である。

- 任意の $0 < \alpha < 1$ に対し、次が成り立つ

$$P[L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})] = 1 - \alpha \quad (16)$$

とき, $[L, U] = [L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})]$ を θ の信頼係数 $1 - \alpha$ の信頼区間 *confidence interval* という。

- L, U をそれぞれ 下側信頼限界 *lower confidence limit* と 上側信頼限界 *upper confidence limit* という。
- $1 - \alpha$ を $[L, U]$ の信頼係数 *confidence coefficient* という。

注: 信頼係数は $1 - \alpha = 0.9, 0.95$ とするのが慣例である。

- 1 標本問題 one sample problem

- ① 1 つの無作為標本:

$$X_1, \dots, X_n \stackrel{i.i.d.}{\sim} N(\mu, \sigma^2)$$

- ② μ の信頼区間の構成
- ③ σ^2 の信頼区間の構成

- 2 標本問題 two sample problem

- ① 2 つの無作為標本:

$$X_1, \dots, X_m \stackrel{i.i.d.}{\sim} N(\mu_x, \sigma_x^2)$$

$$Y_1, \dots, Y_n \stackrel{i.i.d.}{\sim} N(\mu_y, \sigma_y^2)$$

- ② $\mu_x - \mu_y$ の信頼区間の構成
- ③ σ_x^2 / σ_y^2 の信頼区間の構成

μ の信頼区間 (σ^2 が既知)

定理 (正規分布の平均の信頼区間 ; σ^2 が既知)

- ① 無作為標本:

$$X_1, \dots, X_n \stackrel{i.i.d.}{\sim} N(\mu, \sigma^2)$$

- ② σ^2 : 既知

③ $\bar{X} = n^{-1} \sum_{i=1}^n X_i$

このとき、 μ の信頼係数 $1 - 2\alpha$ の信頼区間は次に与えられる

$$P \left[\bar{X} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha} \leq \mu \leq \bar{X} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha} \right] = 1 - 2\alpha$$

ただし、 z_α は標準正規分布 $N(0, 1)$ の α パーセント点である。すなわち

$$\int_{-\infty}^{z_\alpha} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \alpha$$

μ の信頼区間 (σ^2 が既知): 証明

標本平均 \bar{X} は正規分布に従う。また平均と分散はそれぞれ

$$E(\bar{X}) = \mu, V(\bar{X}) = \frac{\sigma^2}{n}$$

すなわち、

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

従って、次が成り立つ

$$Y = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0, 1)$$

$$\begin{aligned} 1 - 2\alpha &= P[|Y| \leq z_{1-\alpha}] \\ &= P\left[-z_{1-\alpha} \leq \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \leq z_{1-\alpha}\right] \\ &= P\left[\bar{X} - \frac{\sigma}{\sqrt{n}}z_{1-\alpha} \leq \mu \leq \bar{X} + \frac{\sigma}{\sqrt{n}}z_{1-\alpha}\right] \end{aligned}$$

μ の信頼区間 (σ^2 が未知)

定理 (正規分布の平均の信頼区間 ; σ^2 が未知)

次の条件が成り立つとする。

- ① t_{α}^{n-1} : 自由度 $n-1$ の t 分布の α パーセント点
- ② 無作為標本:

$$X_1, \dots, X_n \stackrel{i.i.d.}{\sim} N(\mu, \sigma^2)$$

- ③ μ, σ^2 : 未知
- ④ 標本平均, 標本分散

$$\bar{X} = n^{-1} \sum_{i=1}^n X_i, S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

このとき、 μ の信頼係数 $1 - 2\alpha$ の信頼区間は次に与えられる

$$P \left[\bar{X} - \frac{S}{\sqrt{n}} t_{1-\alpha}^{n-1} \leq \mu \leq \bar{X} + \frac{S}{\sqrt{n}} t_{1-\alpha}^{n-1} \right] = 1 - 2\alpha$$

μ の信頼区間 (σ^2 が未知): 証明

- 条件より, まず次が成り立つ。

$$Y = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0, 1)$$

$$Z = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 \sim \chi^2(n-1)$$

- 次に, Y, Z の独立性より, 次が言える。

$$T = \frac{Y}{\sqrt{Z/(n-1)}} = \frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t(n-1)$$

- 以上により, 次の確率的評価が成り立つ。

$$\begin{aligned} 1 - 2\alpha &= P[|T| \leq t_{1-\alpha}^{n-1}] \\ &= P\left[-t_{1-\alpha}^{n-1} \leq \frac{\sqrt{n}(\bar{X} - \mu)}{S} \leq t_{1-\alpha}^{n-1}\right] \\ &= P\left[\bar{X} - \frac{S}{\sqrt{n}}t_{1-\alpha}^{n-1} \leq \mu \leq \bar{X} + \frac{S}{\sqrt{n}}t_{1-\alpha}^{n-1}\right] \end{aligned}$$

μ の信頼区間 (σ^2 が未知): 例

例 (μ の信頼区間; σ^2 が未知)

ある学校で 100 人の生徒が無作為に選ばれ、これらの生徒に知能テストが行われた。テストの点数の平均 50.0, 標本分散 $s_x^2 = 100.0$ が得られた。この学校の生徒の同じテストにおける得点が正規分布 $N(\mu, \sigma^2)$ に従うとして、母平均 μ の信頼係数 $1 - 2\alpha = 95\%$, 90% の信頼区間を求めよ。

解 求める信頼区間は $\left[\bar{x} - \frac{s}{\sqrt{n}} t_{1-\alpha}^{n-1}, \bar{x} + \frac{s}{\sqrt{n}} t_{1-\alpha}^{n-1} \right]$ で、
 $n = 100, \bar{x} = 50.0, \alpha = 0.025, 0.05$ となる。また、

$$s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2} = \sqrt{\frac{n}{n-1} s_x^2} = \sqrt{\frac{100}{99}} \times 10.0 = 10.05$$

表によって、 $t_{1-0.025}^{99} = t_{0.975}^{99} = 1.98$, $t_{1-0.05}^{99} = t_{0.95}^{99} = 1.65$

- 95% 信頼区間: $\left[50.0 - \frac{10.05}{\sqrt{100}} 1.98, 50.0 + \frac{10.05}{\sqrt{100}} 1.98 \right] = [48.0, 52.0]$
- 90% 信頼区間: $\left[50.0 - \frac{10.05}{\sqrt{100}} 1.65, 50.0 + \frac{10.05}{\sqrt{100}} 1.65 \right] = [48.3, 51.7]$

第6講：区間推定：正規分布の分散の場合

正規分布の分散の信頼区間

正規分布の分散 σ^2 の信頼区間の構築について考える。

定理

次の条件が成り立つとする。

- 1 χ_{α}^{n-1} : 自由度 $n-1$ の χ^2 分布の α パーセント点
- 2 無作為標本:

$$X_1, \dots, X_n \stackrel{i.i.d.}{\sim} N(\mu, \sigma^2)$$

3

$$\bar{X} = n^{-1} \sum_{i=1}^n X_i, S^2 = \sum_{i=1}^n (X_i - \bar{X})^2$$

このとき、 μ の信頼係数 $1 - 2\alpha$ の信頼区間は次に与えられる

$$P \left[\frac{S^2}{\chi_{1-\alpha}^{n-1}} \leq \sigma^2 \leq \frac{S^2}{\chi_{\alpha}^{n-1}} \right] = 1 - 2\alpha$$

σ^2 の信頼区間: 証明

まず, 次の事実に注意する。

$$\frac{S^2}{\sigma^2} = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1)$$

従って, $\chi_{1-\alpha}^{n-1}$ を自由度 $n-1$ のカイ自乗分布の α パーセント点とすると, 以下の計算が成り立つ。

$$\begin{aligned} 1 - 2\alpha &= P \left[\chi_{\alpha}^{n-1} \leq \frac{S^2}{\sigma^2} \leq \chi_{1-\alpha}^{n-1} \right] \\ &= P \left[\frac{S^2}{\chi_{1-\alpha}^{n-1}} \leq \sigma^2 \leq \frac{S^2}{\chi_{\alpha}^{n-1}} \right] \end{aligned}$$

Q.E.D.

σ^2 の信頼区間: 例

例 (σ^2 の信頼区間)

ある学校で 100 人の生徒が無作為に選ばれ、これらの生徒に知能テストが行われた。テストの点数の平均 50.0, 標本分散 $s_x^2 = 100.0$ が得られた。この学校の生徒の同じテストにおける得点が正規分布 $N(\mu, \sigma^2)$ に従うとして、母分散 σ^2 の信頼係数 $1 - 2\alpha = 95\%$, 90% の信頼区間を求めよ。

解 求める信頼区間は $\left[\frac{S^2}{\chi_{1-\alpha}^{n-1}}, \frac{S^2}{\sqrt{\chi_{\alpha}^{n-1}}} \right]$ ただし、

$$n = 100, \bar{x} = 50.0, s^2 = \sum (x_i - \bar{x})^2 = ns_x^2 = 100.0^2, \alpha = 0.025, 0.05$$

表によって、

$$\chi_{0.975}^{99} = 128.422, \chi_{0.025}^{99} = 73.361, \chi_{0.95}^{99} = 123.225, \chi_{0.05}^{99} = 77.046$$

- 95% 信頼区間: $[100.0^2/128.422, 100.0^2/73.361] = [77.9, 136.3]$
- 90% 信頼区間: $[100.0^2/123.225, 100.0^2/77.046] = [81.2, 129.8]$

分散の比の信頼区間

定理 (σ_x^2/σ_y^2 の信頼区間)

(A1) $F_{\alpha}^{n-1; m-1}$: 自由度 $(n-1, m-1)$ の F 分布の α パーセント点

(A2) 2つの正規分布からの無作為標本を抽出する。

$$X_1, \dots, X_m \stackrel{i.i.d.}{\sim} N(\mu_x, \sigma_x^2), \quad Y_1, \dots, Y_n \stackrel{i.i.d.}{\sim} N(\mu_y, \sigma_y^2)$$

(A3) $X_1, \dots, X_m; Y_1, \dots, Y_n$: 互いに独立である。

(A4) 次の記号を用いる。

$$\bar{X} = \frac{1}{m} \sum_{i=1}^m X_i, \quad S_x^2 = \frac{1}{m-1} \sum_{i=1}^m (X_i - \bar{X})^2$$

$$\bar{Y} = \frac{1}{n} \sum_{j=1}^n Y_j, \quad S_y^2 = \frac{1}{n-1} \sum_{j=1}^n (Y_j - \bar{Y})^2$$

このとき、 σ_x^2/σ_y^2 の信頼係数 $1 - 2\alpha$ の信頼区間は次に与えられる

$$P \left[\frac{S_x^2}{S_y^2} F_{\alpha}^{n-1; m-1} \leq \frac{\sigma_x^2}{\sigma_y^2} \leq \frac{S_x^2}{S_y^2} F_{1-\alpha}^{n-1; m-1} \right] = 1 - 2\alpha$$

σ_x^2/σ_y^2 の信頼区間: 証明

- 条件より次が成り立つ。

$$(m-1) \frac{S_x^2}{\sigma_x^2} = \frac{1}{\sigma_x^2} \sum_{i=1}^m (X_i - \bar{X})^2 \sim \chi^2(m-1) \quad (17)$$

$$(n-1) \frac{S_y^2}{\sigma_y^2} = \frac{1}{\sigma_y^2} \sum_{j=1}^n (Y_j - \bar{Y})^2 \sim \chi^2(n-1) \quad (18)$$

- また標本の独立性により、次が言える。

$$Z = \frac{S_y^2/\sigma_y^2}{S_x^2/\sigma_x^2} = \frac{S_y^2}{S_x^2} \times \frac{\sigma_x^2}{\sigma_y^2} \sim F(n-1, m-1)$$

- 以上により、次の計算が成り立つ。

$$\begin{aligned} 1 - 2\alpha &= P \left[F_{\alpha}^{n-1; m-1} \leq \frac{S_y^2}{S_x^2} \times \frac{\sigma_x^2}{\sigma_y^2} \leq F_{1-\alpha}^{n-1; m-1} \right] \\ &= P \left[F_{\alpha}^{n-1; m-1} \frac{S_x^2}{S_y^2} \leq \frac{\sigma_x^2}{\sigma_y^2} \leq F_{1-\alpha}^{n-1; m-1} \frac{S_x^2}{S_y^2} \right] \end{aligned}$$

Q.E.D.

平均の差の信頼区間 (分散が既知)

定理 ($\mu_x - \mu_y$ の信頼区間; 分散が既知)

次の条件が成り立つとする。

- ① 無作為標本 (σ_x^2, σ_y^2 : 既知):

$$X_1, \dots, X_m \stackrel{i.i.d.}{\sim} N(\mu_x, \sigma_x^2), \quad Y_1, \dots, Y_n \stackrel{i.i.d.}{\sim} N(\mu_y, \sigma_y^2)$$

- ② $X_1, \dots, X_m; Y_1, \dots, Y_n$ が互いに独立である。
- ③ 次の記号を用意する。

$$\bar{X} = \frac{1}{m} \sum_{i=1}^m X_i, \quad \bar{Y} = \frac{1}{n} \sum_{j=1}^n Y_j,$$

このとき、 $\mu_x - \mu_y$ の信頼係数 $1 - 2\alpha$ の信頼区間は次に与えられる

$$P \left[\bar{X} - \bar{Y} - z_{1-\alpha} \sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}} \leq \mu_x - \mu_y \leq \bar{X} - \bar{Y} + z_{1-\alpha} \sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}} \right] = 1 - 2\alpha$$

$\mu_x - \mu_y$ の信頼区間 (分散が既知): 証明

- $\bar{X} - \bar{Y}$ が正規分布に従い、

$$E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_x - \mu_y$$

$$V(\bar{X} - \bar{Y}) = V(\bar{X}) + V(\bar{Y}) = \frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}$$

- 従って

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\sigma_x^2/m + \sigma_y^2/n}} \sim N(0, 1)$$

- よって、

$$1 - 2\alpha = P[|Z| \leq z_{1-\alpha}]$$

$$= P\left[-z_{1-\alpha} \leq \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\sigma_x^2/m + \sigma_y^2/n}} \leq z_{1-\alpha}\right]$$

したがって

$$1 - 2\alpha = P\left[\bar{X} - \bar{Y} - z_{1-\alpha}\sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}} \leq \mu_x - \mu_y \leq \bar{X} - \bar{Y} + z_{1-\alpha}\sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}\right]$$

平均の差の信頼区間 (分散が未知)

定理 ($\mu_x - \mu_y$ の信頼区間; 分散が未知)

① 無作為標本 (同分散):

$$X_1, \dots, X_m \stackrel{i.i.d.}{\sim} N(\mu_x, \sigma^2), \quad Y_1, \dots, Y_n \stackrel{i.i.d.}{\sim} N(\mu_y, \sigma^2)$$

② $X_1, \dots, X_m, Y_1, \dots, Y_n$ が互いに独立である。

③ 以下の記号を用いる。

$$\bar{X} = \frac{1}{m} \sum_{i=1}^m X_i, \quad S_x^2 = \sum_{i=1}^m (X_i - \bar{X})^2$$

$$\bar{Y} = \frac{1}{n} \sum_{j=1}^n Y_j, \quad S_y^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

$$\hat{\sigma}^2 = \frac{S_x^2 + S_y^2}{m + n - 2}$$

このとき、 $\mu_x - \mu_y$ の信頼係数 $1 - 2\alpha$ の信頼区間は次に与えられる

$$I = \left[\bar{X} - \bar{Y} - t_{1-\alpha}^{m+n-2} \sqrt{\frac{1}{m} + \frac{1}{n}} \hat{\sigma}, \quad \bar{X} - \bar{Y} + t_{1-\alpha}^{m+n-2} \sqrt{\frac{1}{m} + \frac{1}{n}} \hat{\sigma} \right]$$

$\mu_x - \mu_y$ の信頼区間 (分散が未知): 証明

- まず次の事実に注意する。

$$U = \bar{X} - \bar{Y} \sim N\left(\mu_x - \mu_y, \left(\frac{1}{m} + \frac{1}{n}\right)\sigma^2\right)$$

$$V = \frac{S_x^2}{\sigma^2} + \frac{S_y^2}{\sigma^2} \sim \chi^2(m+n-2)$$

- 次に, $W = (U - E(U))/\sqrt{\text{Var}(U)}$ とする。 $\bar{X}, \bar{Y}, S_x^2, S_y^2$ の独立性により, 次が成り立つ。

$$\begin{aligned} T &= W/\sqrt{V/(m+n-2)} \\ &= \frac{[(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)] / \sqrt{(\frac{1}{m} + \frac{1}{n})\sigma^2}}{\sqrt{(S_x^2/\sigma^2 + S_y^2/\sigma^2)/(m+n-2)}} \\ &= \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{1}{m} + \frac{1}{n}} \hat{\sigma}} \sim t(m+n-2) \end{aligned}$$

- 従って, 次の確率的評価が成り立つ。

$$1 - 2\alpha = P \left[\left| \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{1}{m} + \frac{1}{n}} \hat{\sigma}} \right| \leq t_{1-\alpha}^{m+n-2} \right]$$

第7講：仮説検定

仮説検定の基本的概念

Definition (統計的仮説)

statistical hypothesis

- ある確率変数の分布に関する記述を統計的仮説という。
- 統計的仮説が分布の形を完全に指定する場合、この仮説を単純仮説 simple hypothesis といい、それ以外の場合を複合仮説 composite hypothesis という。

Example

X_1, \dots, X_n を $N(\theta, 25)$ からの無作為標本とする。

- 単純仮説 : $H : \theta = 17$
- 複合仮説 : $H : \theta \leq 17$

Definition (仮説検定)

test of a statistical hypothesis

ある統計的仮説を棄却 reject するかどうかを決めるための手順を仮説の検定という。

Example

X_1, \dots, X_n を $N(\theta, 25)$ からの無作為標本とする。

- 仮説: $H: \theta \leq 17$
- 検定法 γ :

$$\bar{X} > 17 + \frac{10}{\sqrt{n}}$$

のとき、 H を棄却する

Definition (標本空間)

sample space

n 個の無作為標本 X_1, \dots, X_n に対して、次の集合を標本空間という。

$$\mathcal{X} = \{(x_1, \dots, x_n); (x_1, \dots, x_n) \text{ は } (X_1, \dots, X_n) \text{ の全ての取りうる値}\}$$

Definition (棄却域)

critical region \mathcal{X} を標本空間とし、 $C \subset \mathcal{X}$ は \mathcal{X} の部分集合で、

$$(x_1, \dots, x_n) \in C \implies$$

のとき、仮説を棄却する。 C をこの検定の棄却域という。

Example

X_1, \dots, X_n を $N(\theta, 25)$ からの無作為標本とする。

- 仮説: $H: \theta \leq 17$
- 検定法 γ : $\bar{X} > 17 + 10/\sqrt{n}$ のとき、 H を棄却する。

2種類の誤り

Definition (帰無仮説・対立仮説)

null hypothesis, alternative hypothesis

- 検定すべき仮説を帰無仮説といい、通常 H_0 で表す。
- 帰無仮説と矛盾する仮説を対立仮説といい、通常 H_1 で表す。帰無仮説と対立仮説は、どちらか一方必ず成り立つとする。

Definition (2種類の誤り)

two types of error

- 帰無仮説 H_0 が成り立つとき、 H_0 を棄却することを第1種の誤り type 1 error といい、第1種の誤りを犯す確率を第1種の誤りの大きさ size of type 1 error という。
- 対立仮説 H_1 が成り立つとき、対立仮説 H_1 を棄却することを第2種の誤り type 2 error といい、第2種の誤りを犯す確率を第2種の誤りの大きさ size of type 2 error という。

検出力関数

Definition (検出力関数)

power function 密度関数 $f(x|\theta)$ を持つ母集団からの無作為標本抽出を考える。
帰無仮説 H_0 に対する検定 γ の検出力関数は

$$\pi_\gamma(\theta) = P [H_0 \text{を棄却} | X_1, \dots, X_n \sim f(x|\theta)]$$

で定義される母数 θ の関数である。 $\pi_\gamma(\theta) \in [0, 1]$ になることに注意する。

Example

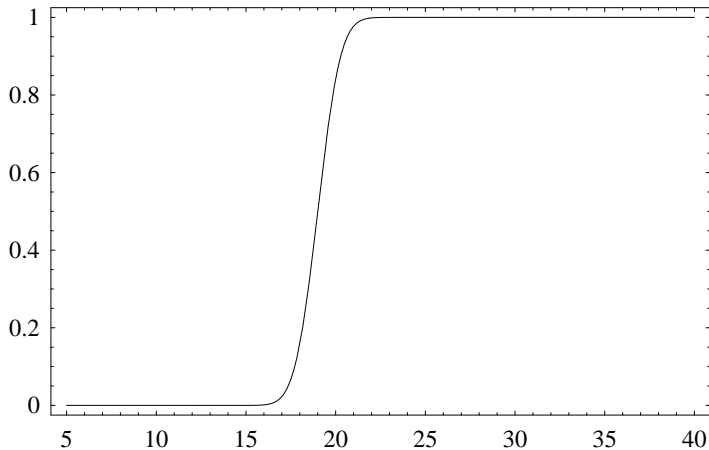
X_1, \dots, X_n を $N(\theta, 25)$ からの無作為標本とする。

- 帰無仮説: $H_0 : \theta \leq 17$
- $\bar{X} > 17 + 10/\sqrt{n}$ のとき、 H_0 を棄却する検定の検出力関数は

$$\begin{aligned}\pi_\gamma(\theta) &= P[\bar{X} > 17 + 10/\sqrt{n} | \theta \in R] \\ &= P\left[\frac{\sqrt{n}(\bar{X} - \theta)}{5} > \frac{17 + 10/\sqrt{n} - \theta}{5/\sqrt{n}}\right] \\ &= 1 - \Phi\left(\frac{17 + 10/\sqrt{n} - \theta}{5/\sqrt{n}}\right)\end{aligned}$$

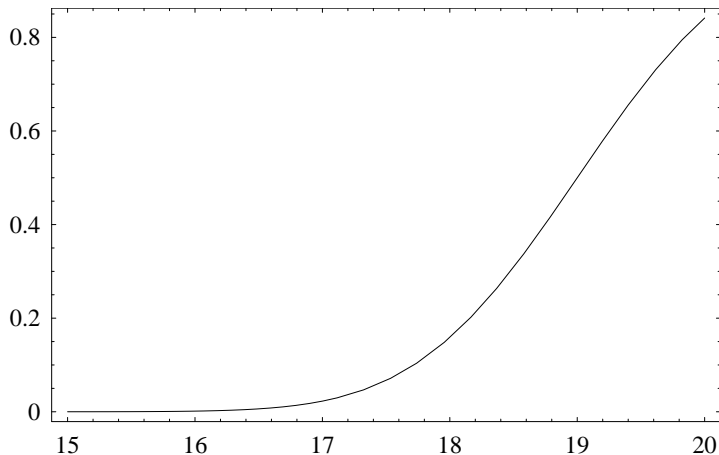
検出力関数 (つづき)

$$\pi_{\gamma}(\theta) = 1 - \Phi\left(\frac{17+10/\sqrt{n}-\theta}{5/\sqrt{n}}\right) \text{ with } n = 25$$



検出力関数 (つづき)

帰無仮説: $H_0 : \theta \leq 17$. 検出力関数 $\pi_\gamma(\theta)$ の $\theta = 17$ の近くの様子



Definition (母数空間)

parameter space 確率変数 X が密度関数 $f(x|\theta)$ に従うとき、母数 θ の全ての取りうる値の集合 Θ を母数空間という。すなわち、 $\Theta = \{\theta \mid \text{全て可能な}\theta\text{の値}\}$

Example

- X が二項分布 $\text{Bi}(n, \theta)$ に従うとする。すなわち

$$P(X = x) = {}_n C_x \theta^n (1 - \theta)^{n-x}$$

このとき、母数空間は $\Theta = \{\theta \mid 0 \leq \theta \leq 1\}$

- X が正規分布 $N(\theta, 1)$ に従うとする。すなわち

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x - \theta)^2}{2}\right\}$$

このとき、母数空間は $\Theta = \{\theta \mid -\infty < \theta < \infty\}$

Definition (有意水準)

size of test (significance level)

- 密度関数: $f(x|\theta)$ からの標本抽出を考える
- 帰無仮説 $H_0 : \theta \in \Theta_0$ に対する検定 γ の検出力関数が $\pi_\gamma(\theta)$ のとき、

$$\text{有意水準} = \sup_{\theta \in \Theta_0} [\pi_\gamma(\theta)]$$

Example

X_1, \dots, X_n を $N(\theta, 25)$ からの無作為標本とする。

- 帰無仮説: $H_0 : \theta \in \Theta_0 = \{\theta \mid \theta \leq 17\}$
- γ : $\bar{X} > 17 + 10/\sqrt{n}$ のとき、 H_0 を棄却する
- 検出力関数: $\pi_\gamma(\theta) = 1 - \Phi\left(\frac{17+10/\sqrt{n}-\theta}{5/\sqrt{n}}\right)$

$$\text{有意水準} = \sup_{\theta \leq 17} \left[1 - \Phi\left(\frac{17 + 10/\sqrt{n} - \theta}{5/\sqrt{n}}\right) \right]$$

第8講：仮説検定：尤度比検定

尤度比検定とは

Definition (尤度比検定)

likelihood ratio test

- $X_1, \dots, X_n \sim f(x|\theta)$
- 帰無仮説 $H_0 : \theta = \theta_0$
- 対立仮説 $H_1 : \theta = \theta_1$
- 尤度比

$$\begin{aligned}\lambda &= \frac{f(X_1|\theta_0) \cdots f(X_n|\theta_0)}{f(X_1|\theta_1) \cdots f(X_n|\theta_1)} \\ &= \frac{\prod_{i=1}^n f(X_i|\theta_0)}{\prod_{i=1}^n f(X_i|\theta_1)}\end{aligned}$$

- 尤度比検定：任意の $k > 0$ に対して、次のように検定する

$$\begin{cases} \lambda \leq k \text{ のとき: } H_0 \text{ を棄却} \\ \lambda > k \text{ のとき: } H_0 \text{ を採択} \end{cases}$$

Definition (最強力検定)

most powerful test

- 帰無仮説 : $H_0 : \theta = \theta_0$; 対立仮説 : $H_1 : \theta = \theta_1$
- 有意水準 α ($0 < \alpha < 1$) : 与えられている
- 検定方式 γ は次が満たされる

$$\pi_\gamma(\theta_0) \leq \alpha$$

次が満たされるとき、検定 γ^* は有意水準 α での最強力検定という。

- 1 $\pi_{\gamma^*}(\theta_0) = \alpha$
- 2 $\pi_{\gamma^*}(\theta_1) \geq \pi_\gamma(\theta_1)$

Theorem (Neyman-Pearson の補題)

Neyman-Pearson Lemma

- $X_1, \dots, X_n \sim f(x|\theta)$
- 帰無仮説 $H_0 : \theta = \theta_0$
- 対立仮説 $H_1 : \theta = \theta_1$
- 尤度比: $\lambda = \prod_{i=1}^n f(X_i|\theta_0) / \prod_{i=1}^n f(X_i|\theta_1)$

有意水準 $\alpha (0 < \alpha < 1)$ での尤度比検定, すなわち

- 1 棄却域: $C^* = \{(x_1, \dots, x_n) \mid \lambda \leq k^*\}$
- 2 有意水準: $\alpha = P[(X_1, \dots, X_n) \in C^* \mid \theta = \theta_0]$

が満たされる尤度比検定 γ^* は有意水準 α での最強力検定となる。

Neyman-Pearson の補題: 証明

- 連続の場合のみについて証明する。
- 任意の有意水準 α 以下の検定 γ を考える。検定 γ の棄却域を C とする。すなわち

$$P[(X_1, \dots, X_n) \in C \mid \theta = \theta_0] \leq \alpha$$

- 次の記号

$$\int_A L_0 = \int \dots \int_A \prod_{i=1}^n f(x_i | \theta_0) dx_1 \cdots dx_n$$

$$\int_A L_1 = \int \dots \int_A \prod_{i=1}^n f(x_i | \theta_1) dx_1 \cdots dx_n$$

を約束すると、 $\pi_{\gamma^*}(\theta_1) = \int_{C^*} L_1$, $\pi_{\gamma}(\theta_1) = \int_C L_1$ となる。従って、証明すべきことは

$$\int_{C^*} L_1 \geq \int_C L_1$$

Neyman-Pearson の補題: 証明 (つづき)

一方

$$\begin{aligned}\int_{C^*} L_1 - \int_C L_1 &= \int_{C^* \cap \bar{C}} L_1 - \int_{C \cap \bar{C}^*} L_1 \\ &\geq \frac{1}{k^*} \int_{C^* \cap \bar{C}} L_0 - \frac{1}{k^*} \int_{C \cap \bar{C}^*} L_0 \\ &= \frac{1}{k^*} \left(\int_{C^* \cap \bar{C}} L_0 + \int_{C^* \cap C} L_0 - \int_{C^* \cap C} L_0 - \int_{C \cap \bar{C}^*} L_0 \right) \\ &= \frac{1}{k^*} \left(\int_{C^*} L_0 - \int_C L_0 \right) \\ &= \frac{1}{k^*} (\alpha - \text{size of } \gamma) \\ &\geq 0\end{aligned}$$

This completes the proof.

尤度比検定：例

- $X_1, \dots, X_n \sim N(\theta, 1)$
- $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$ ($\theta_1 > \theta_0$ を仮定)
- 有意水準を α とする
- 尤度比

$$\begin{aligned}\lambda &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{(X_i - \theta_0)^2}{2}} / \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{(X_i - \theta_1)^2}{2}} \\ &= \exp \left\{ -\frac{1}{2} \sum_{i=1}^n [(X_i - \theta_0)^2 - (X_i - \theta_1)^2] \right\} \\ &= \exp \left[n(\theta_0 - \theta_1) \bar{X} - \frac{n}{2}(\theta_0^2 - \theta_1^2) \right]\end{aligned}$$

従って

$$\begin{aligned}\lambda &\leq k^* \\ \Leftrightarrow \bar{X} &\geq k\end{aligned}$$

尤度比検定：例

最後に、有意水準を達成するように、 k を決める。

$$\begin{aligned}\alpha &= P[\bar{X} \geq k \mid \theta = \theta_0] \\ &= P[\sqrt{n}(\bar{X} - \theta_0) \geq \sqrt{n}(k - \theta_0)] \\ &= 1 - \Phi[\sqrt{n}(k - \theta_0)]\end{aligned}$$

したがって、

$$k = \theta_0 + \frac{1}{\sqrt{n}}\Phi^{-1}(1 - \alpha)$$

すなわち

$$\bar{X} \geq \theta_0 + \frac{1}{\sqrt{n}}\Phi^{-1}(1 - \alpha)$$

のとき、 H_0 を棄却する。

$$\begin{aligned}\alpha = 0.025 & \quad \Phi^{-1}(1 - \alpha) = \Phi^{-1}(0.975) \approx 1.96 \\ \alpha = 0.05 & \quad \Phi^{-1}(1 - \alpha) = \Phi^{-1}(0.95) \approx 1.65 \\ \alpha = 0.10 & \quad \Phi^{-1}(1 - \alpha) = \Phi^{-1}(0.90) \approx 1.28\end{aligned}$$

第9講：仮説検定:信頼区間による方法

仮説検定と信頼区間の双対性

信頼区間による仮説検定の方法を解説する。

- 密度関数 $f(x|\theta)$ をもつ母集団からの標本抽出 X_1, \dots, X_n を考える
- θ の信頼度 $1 - \alpha$ の信頼区間：

$$(L, U) = (L(X_1, \dots, X_n), U(X_1, \dots, X_n))$$

- 帰無仮説 $H_0 : \theta = \theta_0$ に対する検定 γ :
 - $\theta_0 \in (L, U)$ のとき、 H_0 を採択
 - $\theta_0 \notin (L, U)$ のとき、 H_0 を棄却

この検定の有意水準は

$$\begin{aligned} P[H_0 \text{を棄却} | H_0] &= P[\theta_0 \notin (L, U) | H_0] \\ &= 1 - P[\theta_0 \in (L, U) | H_0] \\ &= \alpha \end{aligned}$$

Theorem (平均の信頼区間)

次の条件が成り立つとする。

- ① t_{α}^{n-1} : 自由度 $n-1$ の t 分布の α パーセント点
- ② 無作為標本:

$$X_1, \dots, X_n \stackrel{i.i.d.}{\sim} N(\mu, \sigma^2)$$

- ③ μ, σ^2 : 未知
- ④ 標本平均, 標本分散

$$\bar{X} = n^{-1} \sum_{i=1}^n X_i, S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

このとき、 μ の信頼係数 $1 - 2\alpha$ の信頼区間は次に与えられる

$$P \left[\bar{X} - \frac{S}{\sqrt{n}} t_{1-\alpha}^{n-1} \leq \mu \leq \bar{X} + \frac{S}{\sqrt{n}} t_{1-\alpha}^{n-1} \right] = 1 - 2\alpha$$

正規分布平均の検定 (つづき)

- 帰仮説 $H_0 : \mu = \mu_0$ の検定を考える
- 棄却域 :

$$\begin{aligned} C &= \{(x_1, \dots, x_n); |T| > t_{1-\alpha}^{n-1}\} \\ &= \left\{ (x_1, \dots, x_n); \left| \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} \right| > t_{1-\alpha}^{n-1} \right\} \end{aligned}$$

- 検定の有意水準 :

$$\begin{aligned} \text{有意水準} &= P[H_0 \text{を棄却} | H_0] \\ &= P \left\{ \left| \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} \right| > t_{1-\alpha}^{n-1} \right\} \\ &= 1 - P \left\{ \left| \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} \right| \leq t_{1-\alpha}^{n-1} \right\} \\ &= 1 - (1 - 2\alpha) \\ &= 2\alpha \end{aligned}$$

正規分布平均の検定 (つづき)

ある大学の女子新入生 20 人を抜き取って体重 (単位 : kg) を測定したデータが次のものである。

41 53 48 49 50 55 48 51 45 55 47 56 51 60 55 53 49 66 52 52

女子体重は正規分布に従うとする。このとき有意水準 $\alpha = 0.1, 0.05$ で、女子新入生の体重の平均が 49.5kg であるという仮説を検定せよ。

解 ● $H_0 : \mu = \mu_0 = 49.5$

- $n = 20, \bar{x} = 51.80, s^2 = \sum(x_i - \bar{x})^2 / (n - 1) = 29.22$
- 検定統計量の値 :

$$T = \left| \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} \right| = \left| \frac{\sqrt{20}(51.80 - 50.0)}{\sqrt{29.22}} \right| = 1.90$$

有意水準 = 0.1 $\rightarrow t_{1-0.05}^{n-1} = t_{0.95}^{19} = 1.73 < T \rightarrow H_0$ を棄却

有意水準 = 0.05 $\rightarrow t_{1-0.025}^{n-1} = t_{0.975}^{19} = 2.09 > T \rightarrow H_0$ を採択

正規分布の分散に関する検定

Theorem (分散の信頼区間)

次の条件が成り立つとする。

- ① χ_{α}^{n-1} : 自由度 $n-1$ の χ^2 分布の α パーセント点
- ② 無作為標本:

$$X_1, \dots, X_n \stackrel{i.i.d.}{\sim} N(\mu, \sigma^2)$$

③

$$\bar{X} = n^{-1} \sum_{i=1}^n X_i, S^2 = \sum_{i=1}^n (X_i - \bar{X})^2$$

このとき、 μ の信頼係数 $1 - 2\alpha$ の信頼区間は次に与えられる

$$P \left[\frac{S^2}{\chi_{1-\alpha}^{n-1}} \leq \sigma^2 \leq \frac{S^2}{\chi_{\alpha}^{n-1}} \right] = 1 - 2\alpha$$

正規分布の分散に関する検定 (つづき)

- 帰仮説 $H_0 : \sigma^2 = \sigma_0^2$ の検定を考える
- 棄却域 :

$$C = \{(x_1, \dots, x_n); S^2 < \sigma_0^2 \chi_\alpha^{n-1} \text{ and } S^2 > \sigma_0^2 \chi_{1-\alpha}^{n-1}\}$$

- 検定の有意水準 :

$$\begin{aligned} \text{有意水準} &= P[H_0 \text{を棄却} | H_0] \\ &= P\{S^2 < \sigma_0^2 \chi_\alpha^{n-1} \text{ and } S^2 > \sigma_0^2 \chi_{1-\alpha}^{n-1}\} \\ &= 1 - P\{\sigma_0^2 \chi_\alpha^{n-1} \leq S^2 \leq \sigma_0^2 \chi_{1-\alpha}^{n-1}\} \\ &= 1 - P\left\{\frac{S^2}{\chi_{1-\alpha}^{n-1}} \leq \sigma_0^2 \leq \frac{S^2}{\chi_\alpha^{n-1}}\right\} \\ &= 1 - (1 - 2\alpha) \\ &= 2\alpha \end{aligned}$$

第 10 講：標本数の決定

新薬は本当により有効か？

ある病気に対して、これまでに薬 A が標準的に使われてきた。この薬を患者さんに投与して、ある種の測定値 X を測り、薬 A の有効性を調べることができる。測定値 X が高ければ高いほど、薬が有効であるという。困ったことに、実は、この薬は効き目があまり良くなく、評判がいまひとつである。そこである製薬会社が新薬の開発に乗り出し、数年間を経て、薬 B の開発に成功したという。さて、製薬会社の信念、すなわち

病気の治療に関して、薬 B は薬 A より有効である

という仮説を、統計的にどのように検証したらよいか。この問題に答えるのがこの章の目的である。

統計的判断における標本数の重要性

さて、薬 A, B をそれぞれ n 人の患者に投与し、次の測定値（無作為標本）

$$\text{薬 A} \implies x_1, \dots, x_n$$

$$\text{薬 B} \implies y_1, \dots, y_n$$

を得たとしよう。薬 B が効くか効かないかまだよく分からないこの段階では、倫理的な観点などから、沢山の患者に新薬を投与することはあまり望ましくないであろう。一方、既存薬 A の効き目があまりよくないこともあって、力をいれて開発された新薬による治療を積極的に行いたいのも人情である。すなわち、できれば、少し大きめの標本数 n を取って、治療効果の検討を行いたい。もちろん、標本数 n の決定は、時間や、コスト、施設などの制約を受けることも言うまでもない。

治療効果における標本数の影響

いずれにして、適当な n (例えば $n = 50$) に対して、標本 x_1, \dots, x_n と標本 y_1, \dots, y_n に基づいて、薬 A, B による治療効果の比較を行おう。そのために、それぞれの標本平均

$$\text{薬 A} \iff \bar{x} = \frac{1}{n}(x_1 + \dots + x_n)$$

$$\text{薬 B} \iff \bar{y} = \frac{1}{n}(y_1 + \dots + y_n)$$

を計算し、標本平均の差 $\bar{y} - \bar{x}$ を割り出す。仮に

$$\bar{y} - \bar{x} > 0$$

としよう。測定値が高ければ薬がより効果的なので、この場合、新薬 B は薬 A より治療効果が優れていることを示唆する。しかし、この結論に対する確信の程度は明らかに差 $\bar{y} - \bar{x}$ の大きさに左右される。

標本平均の差が非常に大きい場合

まず $\bar{y} - \bar{x} > 0$ が非常に大きい場合を考える。この場合、自信を持って、「薬 B は薬 A より有効である」と結論付けてよい。

しかし、薬 B の有効性を検出するのに、それほど大きい標本数を必要としなかったかもしれない。時間やコストを無駄にした可能性がある。もちろん当初実験を計画するに当たって、結論をはっきり引き出すために、ある程度大きい標本数を見積もる覚悟があったものの、統計的に、ある数以上多く取った標本を殆ど無駄にした可能性が大いにある。したがって、新薬が有効な場合、

無駄なく、またなるべく新薬が有効と判断するための標本数の決定

という問題が重要となる。

標本平均の差があまり大きくない場合

次に、 $\bar{y} - \bar{x} > 0$ があまり大きくない場合を考える。この場合、 $\bar{y} - \bar{x} > 0$ なので、薬 A がより有効であると信じるよりも、薬 B が優れていると信じたほうがより自然ではあるが、この僅かな差が標本のばらつきによって持たされた偶然の変動によるものと疑われても仕方がない。社会に、大きい声で、「薬 B は薬 A より有効である」とは主張しにくい。

しかしこの程度の患者数に対して、この差 $\bar{y} - \bar{x} > 0$ は、(感覚的に) 僅かとは言え、薬 A と薬 B による真の治療効果を表すのに十分だった可能性もまた残される。したがって、今の場合においては、新薬が有効な場合、

与えられた標本数に対する、統計的証拠の強さ

を客観的に評価する手法が求められる。

以下では、統計的仮説検定の立場から、検出力解析と標本数の決定を中心に、上述の問題を系統的に答えていく。

仮説検定：棄却支持検定の視点から

ここで、 μ_y, μ_x を新薬と標準薬における測定値（＝確率変数）の未知の平均（＝確率変数の期待値）とすると、この平均が高ければ高いほど薬がより有効である。したがって、製薬会社の仮説を統計的に

$$H_1: \theta = \mu_y - \mu_x > 0 \quad (19)$$

と表すことができる。当然ながら、製薬会社は仮説 H_1 が正しいと確信（希望）している。

Definition (研究仮説)

臨床試験や実験などを行い、成立することが望まれる仮説のことを研究仮説といい、通常 H_1 で表す。

したがって、(19) は新薬の開発における製薬メーカーの研究仮説となる。

Definition (対立仮説)

研究仮説 H_1 を、（普通の教科書では）、対立仮説ともいう。

Definition (帰無仮説)

帰無仮説は対立仮説 H_1 と論理的に背反な仮説 H_0 で、また H_1 と H_0 の一方が必

棄却支持検定における有意水準と検出力の役割

現実的に、帰無仮説 H_0 は真か偽かのいずれかであるので、仮説検定において、データを取る前に次の表で表されている状況が考えられる。

Table: 棄却支持検定における2種類の過誤

		真の状況	
		H_0	H_1
決定	H_0	正しい受容	第2種の過誤： $\beta = 1 - \Pr(H_0 \text{を棄却} H_1)$
	H_1	第1種の過誤： $\alpha = \Pr(H_0 \text{を棄却} H_0)$	正しい棄却

2種類の誤り

- α (第1種の過誤)と β (第2種の過誤)は誤りなので、客観的には、当然両方とも小さいほうが望ましい。
- しかし、棄却支持検定において、第1種の過誤を犯すと、研究仮説 H_1 が支持されるので、研究者は第1種の過誤には気にしない。むしろ、第1種の過誤を大きく設定したいところである。反対の立場にある当局や雑誌の編集者は、なるべく第1種の過誤 α を小さくしたい。妥協した結果、仮説検定において、通常 $\alpha = 5\%$ となるように検定の手続きを導く。

Definition

第1種の過誤 α を、検定の大きさ、または検定の有意水準ともいう。

- 一方、棄却支持検定において、第2種の過誤を犯すと、不幸にも正しい研究仮説 H_1 が否定されることになる。したがって、研究者はなるべく第2種の過誤 $\beta = 1 - \Pr(H_0 \text{を棄却} | H_1)$ を小さくしたい。すなわち、 $1 - \beta = \Pr(H_0 \text{を棄却} | H_1)$ をなるべく大きくしたい。

Definition

研究仮説が正しいときに、帰無仮説を棄却する確率、すなわち $1 - \beta = \Pr(H_0 \text{を棄却} | H_1)$ を検出力という。

棄却支持検定の要点

- 研究者は帰無仮説 H_0 を棄却したい.
- 当局は第 1 種の過誤をコントロールしたい.
- 研究者は第 2 種の過誤に神経質である.
- 標本数が大きいければ, 既定の検出力を達成でき, 研究者に有利である.
- 標本数がどうしても不足な場合, 検出力を達成するために, 研究者は当局に検定の有意水準の緩和を要請することが考えられる.

総合すると, 棄却支持検定において, 通常, 研究者は検定の有意水準 $\alpha = 5\%$ を守りながら, なるべく検出力の高い (通常 0.8 以上) 検定を採用したい.

t-検定による標本数の決定 (1 標本片側検定)

X_1, X_2, \dots, X_n を, 平均 μ と分散 σ^2 が共に未知の正規分布 $N(\mu, \sigma^2)$ からの無作為標本とする. ここで, 関心の対象は平均 μ にあり, 分散 σ^2 は局外 (攪乱) 母数といわれることもある. この局外母数である分散 σ^2 を次の標本不偏分散

$$S^2 = \frac{1}{n-1} [(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_n - \bar{X})^2]$$

で推定することが通常よく行われる. ただし,

$$\bar{X} = \frac{1}{n} (X_1 + X_2 + \dots + X_n)$$

は標本平均を表していて, \bar{X} もやはり平均 μ の不偏推定量である. ここで, 次の帰無仮説と研究仮説

$$H_0 : \mu = 0, \quad H_1 : \mu > 0$$

を考える. 有意水準 (第 1 種の過誤の大きさ) α は既に与えられているとする.

t -検定による標本数の決定 (つづき)

このときに, $T = \sqrt{n}(\bar{X} - \mu)/S$ が自由度 $n - 1$ の t 分布に従うことが知られている. この事実を利用すれば, 次の検定

$$\frac{\sqrt{n}\bar{X}}{S} \geq t_{1-\alpha}^{n-1} \implies H_0 \text{ を棄却} \quad (20)$$

の有意水準が α となることが確かめられる. ただし, (20) において, $t_{1-\alpha}^{n-1}$ は自由度 $n - 1$ の t 分布の上側 $1 - \alpha$ パーセント点である.

t distribution with $d.f. = n - 1$

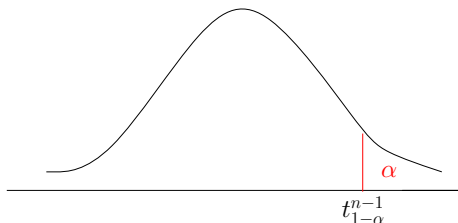


Figure: 片側検定

t -検定による標本数の決定 (つづき)

さらに, 検定 (20) の検出力が

$$\begin{aligned}\text{検出力} &= \Pr \left\{ \frac{\sqrt{n}(\bar{X} - \mu)}{S} \geq t_{1-\alpha}^{n-1} - \frac{\sqrt{n}\mu}{S} \mid H_1 \right\} \\ &= 1 - \Pr \left\{ \frac{\sqrt{n}(\bar{X} - \mu)}{S} \leq t_{1-\alpha}^{n-1} - \frac{\sqrt{n}\mu}{S} \mid H_1 \right\}\end{aligned}$$

と計算できる. したがって, 検出力が $1 - \beta$ となるためには, 標本数 n は次の式

$$t_{1-\alpha}^{n-1} - \frac{\sqrt{n}\mu}{S} = t_{\beta}^{n-1}$$

を満たす必要がある. $t_{1-\alpha}^{n-1} = -t_{\alpha}^{n-1}$ に注意すると, 上の式は次のように

$$n = \frac{S^2}{\mu^2} \left(t_{\alpha}^{n-1} + t_{\beta}^{n-1} \right)^2 \quad (21)$$

と整理される. ただし, μ は対立仮説 (研究仮説) における値である. またこの方程式の両辺に n に依存していることに注意してほしい.

(21) により, 標本数 n は標本分散に比例して増大し, また帰無仮説からの乖離の度合いを表す μ の 2 乗に比例して減少する. $CV = S/\mu$ は変動係数と呼ばれる量である. すなわち, 標本数 n は変動係数の 2 乗に比例して増加する.

ソフトウェア R の利用

さて n に関する方程式 (21) を R を利用して求めてみよう. R では `power.t.test` という関数が用意されている. 次のように入力してみよう.

```
power.t.test(power = 0.8, delta = 1.0, sd = 1.5,  
             sig.level = 0.05, type = "one.sample", alternative = "one.sided")
```

ここで, `power` には目標検出力の値を, `delta` には μ の推定量である標本平均 \bar{X} の値を, `sd` には S の値をそれぞれ代入する. `sig.level = 0.05` は有意水準が 5% であるという意味で, 通常これを弄ることはない. 最後の, `type = "one.sample", alternative = "one.sided"` は, 1 標本問題における片側検定を考えているという意味です. 出力は次のようになる.

```
One-sample t test power calculation
```

```
      n = 15.35763  
delta = 1  
      sd = 1.5  
sig.level = 0.05  
      power = 0.8  
alternative = one.sided
```

ソフトウェア R の利用 (つづき)

なので、今の場合、検出力が 0.8 以上になるためには、標本数 $n \geq 16$ を取る必要がある。 n の情報のみが知りたいときには、次のようにすればよい。

入力

```
samplesize <- power.t.test(power = 0.8, delta = 1.0, sd = 1.5,  
  sig.level = 0.05, type = "one.sample", alternative = "one.sided")  
samplesize$n
```

出力

```
[1] 15.35763
```

目標検出力 $1 - \beta$ の値を色々に変化させて、 n の値を調べるには、次のようにすればよい。

```
result <- 0  
the.power <- 0.8; k <- 10  
for(i in 1:k){  
  samplesize <- power.t.test(power = the.power, delta = 1.0, sd = 1.5,  
    sig.level = 0.05, type = "one.sample", alternative = "one.sided")  
  result[i] <- samplesize$n  
  the.power <- the.power+0.02  
}
```

ソフトウェア R の利用 (つづき)

ここでは, $1 - \beta = 0.8$ から 0.02 の刻みで 0.98 まで増えたときに, n の様子を調べたプログラムである. 次がその結果である.

[1] 15.36 16.19 17.11 18.15 19.32 20.70 22.36 24.46 27.36 32.19

このように検出力の増加の代価を事前に見積ることができる.

この節の最後に, (21) に基づく簡便式を与えよう. そのために, 自由度 $n - 1$ の t 分布を標準正規分布で近似し, パーセント点 t_{α}^{n-1} を標準正規分布の対応するパーセント点 z_{α} で置き換えれば,

$$n = \frac{S^2}{\mu^2} (z_{\alpha} + z_{\beta})^2$$

が得られる. ただし, $z_{\alpha} < t_{\alpha}^{n-1}$ という事実に注意すると, 上の公式による標本数が (21) 式のそれに比べると, 常に小さくなっていることに注意が必要である.

両側検定 vs. 片側検定

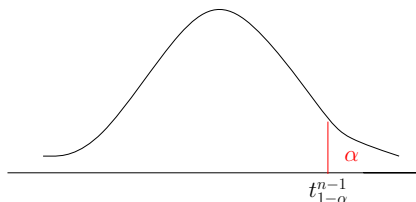
両側検定は次のような検定

$$H_0 : \mu = 0, \quad H_1 : \mu \neq 0$$

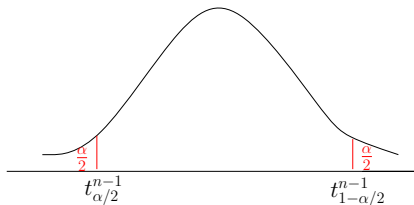
のことをいい、方向性のない検定ともよばれる。このとき、帰無仮説を棄却して対立仮説 $H_1 : \mu \neq 0$ を受容しても、 $\mu > 0$ なのか、それとも $\mu < 0$ なのかは判断を行わない。

両側検定は、片側検定に比べて、一定の有意水準のもとで、検定統計量の値が帰無分布の更なる裾の部分に入らないと、帰無仮説は棄却されない。これは、有意水準を、研究者にとってより厳しく（小さく）設定していることを意味する。有意水準が低くなると、帰無仮説は棄却されにくくなり、検出力が応じて低下する。したがって、検出力を維持するためには、標本数は片側検定に比べて高く設定する必要がある。

t distribution with d.f. = n - 1



t distribution with d.f. = n - 1



両側検定 vs. 片側検定 (つづき)

理論は省略するが、R を用いた計算法のみを示す。次が入力例である。

```
power.t.test(power = 0.8, delta = 1.0, sd = 1.5,  
             sig.level = 0.05, type = "one.sample", alternative = "two.sided")
```

上の例において、片側検定との違いは、`alternative = "one.sided"` を `alternative = "two.sided"` に変えただけである。この例では、 $n \approx 19.67$ という結果になった。片側検定の時には、 $n \approx 15.36$ だったので、標本数が増えていることがわかる。

ここで、両側検定を用いて、 $1 - \beta = 0.8$ から 0.02 の刻みで 0.98 まで増えたときに、 n の値を求め、また片側検定の場合と比較してみよう。プログラムは次の通りである。

```
n.onesided <- 0; n.twosided <- 0  
the.power <- 0.8; k <- 10  
for(i in 1:k){  
  samplesize.one <- power.t.test(power = the.power, delta = 1.0, sd =  
    sig.level = 0.05, type = "one.sample", alternative = "one.sided")  
  n.onesided[i] <- samplesize.one$n  
  samplesize.two <- power.t.test(power = the.power, delta = 1.0, sd =
```

両側検定 vs. 片側検定 (つづき)

得られた結果は次の表に纏めた.

Table: 片側検定と両側検定による標本数の差 : 1 標本の場合

	検出力	n (片側)	n (両側)	差
1	0.80	15.36	19.67	4.31
2	0.82	16.19	20.61	4.42
3	0.84	17.11	21.64	4.53
4	0.86	18.15	22.80	4.65
5	0.88	19.32	24.11	4.79
6	0.90	20.70	25.64	4.94
7	0.92	22.36	27.47	5.12
8	0.94	24.46	29.79	5.33
9	0.96	27.36	32.97	5.61
10	0.98	32.19	38.23	6.04

2 標本問題の計算例

実際に最もよく現れるのがこの場合かもしれない。ここでは、片側検定の場合 (alternative = "one.sided") のみについて解説する。2 標本問題においては、対応がある場合と対応がない場合の2つの場合に分かれる。対応がある場合の検定は、本質的に1 標本問題に帰着でき、直感的にも対応がない場合に比べて、検出力が高いことが想像できる。

対応がない場合に、2つの母集団が同じ分散をもつことを仮定する。すなわち、次を仮定する。

$$X_1, \dots, X_m \stackrel{i.i.d.}{\sim} N(\mu_x, \sigma^2), \quad Y_1, \dots, Y_n \stackrel{i.i.d.}{\sim} N(\mu_y, \sigma^2)$$

このとき、 σ^2 は次のように推定する。

$$\begin{aligned}\bar{X} &= \frac{1}{m} \sum_{i=1}^m X_i, \quad S_x^2 = \sum_{i=1}^m (X_i - \bar{X})^2 \\ \bar{Y} &= \frac{1}{n} \sum_{j=1}^n Y_j, \quad S_y^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2 \\ \hat{\sigma}^2 &= \frac{S_x^2 + S_y^2}{m + n - 2}\end{aligned}$$

さて、R で計算するために、次のように入力すればよい。

2 標本問題の計算例 (つづき)

```
power.t.test(power = 0.8, delta = 1.0, sd = 1.5, sig.level = 0.05,  
             type = "two.sample", alternative = "one.sided")
```

ここで, `type = "two.sample"`は対応がない場合の2標本検定を意味する. また `sd` は上述の $\hat{\sigma}$ の値を代入する. 今回の場合の答えは, $n = 28.52$ で, これがそれぞれの標本セットに含まれるデータの数である. したがって, 実際の標本数はこの2倍となる.

一方, `type = "paired"`にすれば, 対応がある場合の標本数の計算となる.

```
power.t.test(power = 0.8, delta = 1.0, sd = 1.5, sig.level = 0.05,  
             type = "paired", alternative = "one.sided")
```

この場合の答えは対 (ペア) の数となり, 総数はやはりこの2倍となる. この場合, n 組のペアがあり, それぞれ差を作って, `sd` は, この1標本のデータに対する標本標準偏差である. 上の答えは, $n = 15.36$ である.

第 11 講：生存時間解析入門

生存時間解析とは

表 15-1 では、ある治療法（群 1）とプラシーボ（群 2）を、それぞれ 21 人の白血病患者に対して行い、治療開始から死亡するまでの時間（生存週数）を記録したものである。たとえば、新しい治療を受けてから、6 番目の患者は 23 週間後死亡し、またプラシーボを受けた 1 番目の患者も同じ 23 週間後死亡した。生存時間に“+”がついているものは、打切り censoring を意味する。すなわち、研究が終了した時点で、死亡が観察されないか、もしくは試験の途中で脱落したなどを意味する。たとえば、群 1 の 1 番目の患者の生存時間は 35+ なので、この患者は少なくとも 35 週間生存したことを意味する。

表 15-1 白血病患者の週生存時間 (出典：Kleinbaum, 1996, p.75)

患者 番号	生存 時間	白血球数 の対数	性 別	患者 番号	生存 時間	白血球数 の対数	性 別
1	35+	1.45	1	1	23	1.97	1
2	34+	1.47	1	2	22	2.73	0
3	32+	2.20	1	3	17	2.95	0
4	32+	2.53	1	4	15	2.30	0
5	25+	1.78	1	5	12	1.50	0
6	23	2.57	1	6	12	3.06	0
7	22	2.32	1	7	11	3.49	0

生存時間解析とは (つづき)

このように、ある時点から、興味のあるイベント event が発生するまで、個体を観察することが、医学や工学などの分野でしばしば行われる。白血病の場合のイベントは患者の死亡であり、また、たとえば、製品の信頼性実験の場合、製品の故障・破壊をイベントとすることが多い。イベントは、故障 failure や死亡などとも呼ばれる。イベントが観察されるまでの時間 T を生存時間 survival time といい、 T は確率変数である。

生存時間解析の目的は、表 15-1 で示されたような生存時間と関連情報についてのデータを用いて、生存率の推定や群 2 の生存率の比較、さらに生存率と共変量の関係（たとえば白血球数と生存時間の関係）の解明などである。

生存時間解析の主要な目的の 1 つは、次の生存関数 survival function

$$S(t) = \Pr(T > t) = \int_t^{\infty} f(x) dx \quad (22)$$

の推定や比較である。ここで、 $f(t)$ は T の密度関数である。生存関数 $S(t)$ は故障・死亡するまでの時間が t を超える確率を表している。

ところで, T の分布に対するモデルを構築するとき, 次のハザード関数 hazard function

$$h(t) = \lim_{\Delta t \rightarrow 0^+} \frac{\Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \quad (23)$$

を使用するのが便利である. ハザード関数 $h(t)$ は, t 時まで生存した条件の下で, 次の時刻に死亡する, 瞬間死亡率を表している. 人口データなどの解析において, ハザード関数 $h(t)$ の定性的な性質についてある程度経験的に知られていることが多い. $h(t)$ と $S(t)$ は

$$S(t) = \exp \left\{ - \int_0^t h(u) du \right\}, \quad h(t) = - \frac{S'(t)}{S(t)} \quad (24)$$

という関係があり, 1つが決まればもう1つの方を求めることができる.

様々なハザード関数

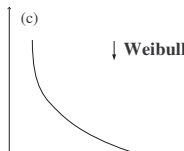
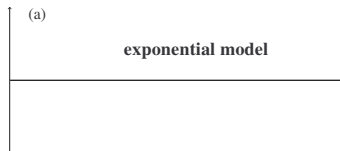
もっとも良く使われるモデルは次のものである。

(i) 指数分布 これはハザードが一定のモデルで (図 15-1(a)) , 観察期間中に健康状態が安定な人を観察するときなどに相当する. 定数 $\lambda > 0$ に対して, $h(t) = \lambda$ とすれば, (24) より生存関数と密度関数が次のように得られる.

$$S(t) = e^{-\lambda t}, \quad f(t) = \lambda e^{-\lambda t} \quad (25)$$

(ii) ワイブル分布 これは指数分布の一般化で, ハザード関数が $h(t) = \lambda p(\lambda t)^{p-1}$ と表され, $p = 1$ のとき $h(t)$ は定数で, $p > 1 (< 1)$ のとき $h(t)$ は t の単調増加 (減少) 関数となる. 図 15-1(b)(c) 参照. たとえば, ある病気に対して, まったく治療を受けてない患者の死亡するハザードは時間と共に増加し, また手術などを受けた患者に対してはハザードが単調減少すると考えられる. (24) より, 生存関数と密度関数は次のようになる.

$$S(t) = e^{-(\lambda t)^p}, \quad f(t) = \lambda p(\lambda t)^{p-1} e^{-(\lambda t)^p} \quad (26)$$



生存関数の最尤推定量

無作為標本 t_1, \dots, t_n に打ち切りがなく、またパラメトリック・モデルの仮定が妥当な場合を考える。生存関数を $S(t) = S(t|\theta)$ と書くと、母数 θ の最尤推定量を $\hat{\theta}$ を用いて、生存関数を $\hat{S}(t) = S(t|\hat{\theta})$ で推定できる。

たとえば、指数分布モデルに対して、 $\bar{T}_n = n^{-1} \sum_{i=1}^n t_i$ を標本平均とすると、 λ の最尤推定量は $\hat{\lambda} = 1/\bar{T}_n$ となる。したがって生存関数の最尤推定量は $\hat{S}(t) = e^{-\hat{\lambda}t}$ となる。表 15-1 の群 2 のデータに対して指数分布モデルを適用してみよう。このとき $\bar{T}_n = 8.667$ で、 $\hat{\lambda} = 0.115$ となる。生存関数のグラフは図 15-2 で示されている。

カプラン・マイヤー推定量

前節で紹介したパラメトリック・モデルの適用が難しい場合、生存関数の推定量を次のように構成することができる。

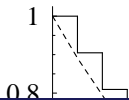
まず、打ち切りがない場合について考える。

まず n 個の無作為標本に対して、打ち切りがない場合を考える。生存関数と分布関数の関係 $S(t) = 1 - F(t)$ により、経験分布 $F_n(t)$ を用いて、 $S(t)$ を

$$\hat{S}(t) = 1 - F_n(t) = \frac{1}{n} \sum_{i=1}^n \delta(t_i \geq t) \quad (28)$$

で推定することが考えられる。ここで $\delta(t_i \geq t)$ は指標関数で、 $t_i \geq t$ のときに 1、 $t_i < t$ のときに 0 である。(28) 式より、 $t \leq t_1$ に対して $\hat{S}(t) \equiv 1$ で、また $t > t_k$ であれば $\hat{S}(t) \equiv 0$ である。推定量 $\hat{S}(t)$ は後述で述べるカプラン・マイヤー推定量の特殊な場合である。

(28) 式を表 15-1 の群 2 のデータに対して適用し得られた生存関数の推定量が図 15-2 である。このように、タイ tie がなければ、 $\hat{S}(t)$ は死亡時刻毎に $1/n$ ずつ減少する階段関数である。



打ち切りがある場合

いま n 個のデータに打ち切りの可能性がある場合を考える. 死亡があった時刻を

$$t_1 < t_2 < \cdots < t_k, \quad k \leq n$$

とする. タイ或いは打ち切りがある場合, $k < n$. ここで時刻 t_j における死亡数を d_j とすると, $D = \sum_{j=1}^n d_j$ は総死亡数を表し, 打ち切りがなければ, $n = D$ で, そうでなければ $n < D$ となる.

次に打ち切り標本数を考える. 区間 $[t_j, t_{j+1})$ における打ち切り標本数を m_j , $j = 1, \dots, k$ とし, 時間 t_j まで (t_j を含まない) の生存者数を n_j とすると, n_j は時刻 t_j におけるリスク集合の大きさと呼ばれる. 次が成り立つことに注意する.

$$n_j = \sum_{i=j}^k (d_i + m_i), \quad i = j, \dots, k$$

生存関数 $S(t)$ のカプラン・マイヤー推定量 (積極限推定量) は次のように定義される.

$$\hat{S}(t) = \begin{cases} 1 & t < t_1 \text{ のとき} \\ \prod_{t_i \leq t} \frac{n_i - d_i}{n_i} & t \geq t_1 \text{ のとき} \end{cases} \quad (29)$$

すなわち, $t_\ell \leq t < t_{\ell+1}$ のとき, 生存関数は次のように計算される.

$$\hat{S}(t) = \frac{n_1 - d_1}{n_1} \cdots \frac{n_\ell - d_\ell}{n_\ell}$$

打ち切りがある場合 (つづき)

打ち切りがまったくない標本に対しては, $n_j = \sum_{i=j}^k d_i$ より, $n_j - d_j = n_{j+1}$ となる. したがって, $t_\ell \leq t < t_{\ell+1}$ に対して, (29) より次が成り立つ.

$$\hat{S}(t) = \frac{n_2}{n_1} \times \frac{n_3}{n_2} \times \cdots \times \frac{n_{\ell+1}}{n_\ell} = \frac{n_{\ell+1}}{n_1}$$

$n_1 = n$ に注意すると, 上の式は (28) とまったく同じものとなることが分かる. 故に, (28) 式は Kaplan-Meier 推定量の特殊な場合に過ぎない.

時刻 t_j における Kaplan-Meier 推定量は, 直前の時刻 t_{j-1} における推定量と, t_j まで生きていた条件のもとでの t_j を乗り越える確率との積で表すことができる. すなわち

$$\hat{S}(t_j) = \hat{S}(t_{j-1}) \times \hat{\Pr} [T > t_j | T \geq t_j]$$

という関係が成立する. $S(t_{j-1}), S(t_{j-2}), \dots, S(t_1)$ についても同様な式を当てはめると, 次が得られる

$$\hat{S}(t_j) = \prod_{i=1}^j \hat{\Pr} [T > t_i | T \geq t_i] \quad (30)$$

このように Kaplan-Meier 推定量は, 条件付生存確率の積として表現できることがわかる.

打ち切りがある場合 (つづき)

カプラン・マイヤー推定量 (29) を表 15-1 の白血病患者データに対して適用し、各死亡時刻における生存関数の推定値を示したのが表 15-2 である。

表 15-2 白血病患者データに対するカプラン・マイヤー推定量

	t_j	d_j	m_j	n_j	$\hat{S}(t_j)$		t_j	d_j	m_j	n_j	$\hat{S}(t_j)$
群 1	6	3	1	21	0.857	群 2	1	2	0	21	0.905
	7	1	1	17	0.807		2	2	0	19	0.810
	10	1	2	15	0.753		3	1	0	17	0.762
	13	1	0	12	0.690		4	2	0	16	0.667
	16	1	3	11	0.627		5	2	0	14	0.571
	22	1	0	7	0.538		8	4	0	12	0.381
	23	1	5	6	0.448		11	2	0	8	0.286
							12	2	0	6	0.190
							15	1	0	4	0.143
							17	1	0	3	0.095
							22	1	0	2	0.048
							23	1	0	1	0.000

打ち切りがある場合（つづき）

表 15-2 の計算結果に基づいて、2 群の生存関数のカプラン・マイヤー推定量を示したのが図 15-3 である。すべての時間において、群 1 の生存関数の推定量が群 2 のそれと比べると明らかに高くなっていることから、治療効果があることが伺える。

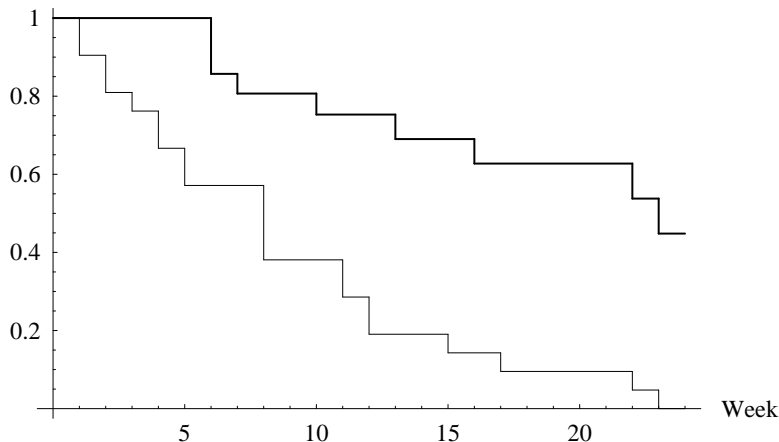


図 15-3 から治療群における生存確率がより高い可能性を示唆しているが、ここで 2 つの生存関数の差の有無の検定について考えてみよう。そのために、2 群における死亡のあったすべての時刻を、 $t_1 < t_2 < \dots < t_K$ とする。第 1 群の時刻 t_j における死亡数、打ち切り数、およびリスク集合の大きさを、それぞれ d_{1j} , m_{1j} , n_{1j} とし、第 2 群対応する量を d_{2j} , m_{2j} , n_{2j} とする。白血病患者データに対してこのように整理したのが表 15-3 である。

このとき、 $d_{1j} + d_{2j}$ は時刻 t_j における 2 群の総死亡数を表し、 $n_{1j} + n_{2j}$ は t_j における総リスク集合の大きさを意味する。2 つの生存曲線に差がなければ、 t_j 時におけるリスク集合の相対的大きさ $n_{1j}/(n_{1j} + n_{2j})$ を用いて、時刻 t_j における群 1 と第 2 群の期待される死亡数を、それぞれ次のように表すことができよう。

$$D_{1j} = \frac{n_{1j}}{n_{1j} + n_{2j}} (d_{1j} + d_{2j}), \quad D_{2j} = \frac{n_{2j}}{n_{1j} + n_{2j}} (d_{1j} + d_{2j}), \quad j = 1, \dots, K$$

例：白血病データ

表 15-3 すべての白血病患者における死亡時刻, 打ち切り数とリスク集合

t_j	群 1			群 2		
	d_{1j}	m_{1j}	n_{1j}	d_{2j}	m_{2j}	n_{2j}
1	0	0	21	2	0	21
2	0	0	21	2	0	19
3	0	0	21	1	0	17
4	0	0	21	2	0	16
5	0	0	21	2	0	14
6	3	1	21	0	0	12
7	1	0	17	0	0	12
8	0	1	16	4	0	12
10	1	1	15	0	0	8
11	0	1	13	2	0	8
12	0	0	12	2	0	6
13	1	0	12	0	0	4
15	0	0	11	1	0	4
16	1	0	11	0	0	3
17	0	3	10	1	0	3

例：白血病（つづき）

ここでそれぞれの群における観測死亡数と期待死亡数を、すべての死亡時刻に対して和をとると次のようになる。

$$O_1 = \sum_{j=1}^K (d_{1j} - D_{1j}), \quad O_2 = \sum_{j=1}^K (d_{2j} - D_{2j})$$

もし両群の生存関数にあまり差がなければ、 O_1 も O_2 も大きくなると想像できよう。一方、簡単な計算より、 $O_1 = -O_2$ となることが分かる。

さて、 $O_1 = -O_2$ は漸近的に平均ゼロの正規分布に従うことが知られ、また O_1 と O_2 の分散を次の式で推定することができる。

$$\hat{V} = \sum_{j=1}^K (d_{1j} + d_{2j}) \frac{n_{1j}}{n_{1j} + n_{2j}} \left(1 - \frac{n_{1j}}{n_{1j} + n_{2j}} \right) \frac{n_{1j} + n_{2j} - d_{1j} - d_{2j}}{n_{1j} + n_{2j} - 1}$$

このように、2群の生存関数に差がないという帰無仮説に対して、次の統計量

$$\chi^2 = \frac{O_1^2}{\hat{V}} = \frac{O_2^2}{\hat{V}} \quad (33)$$

を利用することができる。 χ^2 をログ・ランク検定統計量といい、帰無仮説のもとで、 χ^2 は漸近的に自由度1のカイ二乗分布に従う。

表15.3の白血病データに基づいて、計算してみると

比例ハザード・モデル

ところで、白血病患者の生存時間を示した表 15-1 には、患者の白血球数（の対数）などの情報も示されている。特に白血球数は、白血病患者の死亡をイベントとした場合、よく知られる予後因子 prognostic indicator である。すなわち白血病患者の生存時間は、治療効果のほか、白血球数という予後因子によって影響される可能性が考えられる。このようにいくつかの重要な共変量が存在するとき、それらを解析に入れ、交絡要因による影響を排除して、生存関数の比較などを行うことが重要である。

一般に患者に付随する共変量を $x = (x_1, \dots, x_p)$ とする。ここで x は時間に依存しないことを仮定する。白血病の例の場合、共変量を $x = (x_1, x_2)$ とする。ここで x_1 はダミー変数で、治療を受けた場合、 $x_1 = 0$ 、対象群の場合、 $x_1 = 1$ とし、また x_2 を白血球数の対数とする。この場合、 x_2 の影響を除いて、 x_1 の効果を調べるのが目的である。

さて従来の回帰分析の考え方を借りて、生存時間解析における回帰分析の考え方を述べよう。従来の回帰分析においては、 $E(Y) = g(\beta'x)$ などの仮定をおき、最小二乗法や最尤法などを用いて回帰母数 β の推定を行う。いまの場合、 $E(Y)$ の代わりに、ハザード関数を用いて考えるのが自然であろう。

たとえば、ハザードが時間に依存しなければ、次の単純なモデルが考えられる。

$$h(t; x) = \exp(\alpha + \beta'x) = h_0 e^{\beta'x}, \quad h_0 > 0 \quad (34)$$

モデル (34) におけるハザードは時間と無関係なので、生存時間 T の分布は指数

比例ハザードモデル

指数回帰モデル (34) における h_0 を $h_0(t)$ で置き換えて得られたのが、次のコックス D.R. Cox による比例ハザードモデル proportional hazard model である。

$$h(t; x) = h_0(t) e^{\beta' x}, \quad h_0(t) > 0 \quad (35)$$

ここで $h(t; x = 0) = h_0(t)$ となることから、 $h_0(t)$ を基準ハザード baseline hazard と呼ばれる。 $h_0(t)$ は t の関数であるが、その形を全く指定しない。一方、(35) における共変量の効果については、明示的な関数を用いて規定している。このことから、比例ハザードモデルはセミ・パラメトリックなモデルと呼ばれる。比例ハザードモデルは理論と応用の両面において極めて重要なモデルである。このモデルの一番の魅力は、基準ハザード関数 $h_0(t)$ の形については、如何なるモデルも仮定しない、また仮定する必要がないことである。このことは、生存時間 T の分布がどんな分布であっても、比例ハザードモデル (35) に基づいて解析を行えば、得られる結果は頑健的 robust であることを意味する。基準ハザード関数 $h_0(t)$ は、共変量を考慮しないときのハザード関数であり、実際のデータ解析においてこれを指定する必要がないことは非常に有難いことである。

比例ハザードモデル (つづき)

比例ハザードモデルを適用するときの最大の注意点は比例ハザード性の仮定といえよう. いま共変量 x, x^* を持つ二人のハザードの比を考えると, (35) より

$$\frac{h(t; x)}{h(t; x^*)} = \exp [\beta'(x - x^*)] \quad (36)$$

となる. すなわち個人間のハザード比 hazard ratio は, 時間によらず, 共変量のみ関数となる. ハザード関数における (36) 式の制約は通常比例ハザード性 proportional hazard assumption と呼ばれる. 共変量の効果が時間と共に変化するなどの場合に, 比例ハザード性は成り立たず, 吟味せずに比例ハザードモデルを当てはめることは不適切である.

ところで, 比例ハザードモデル (35) のもとで, 生存関数は (24) より次のように表現できる.

$$\begin{aligned} S(t; x) &= \exp \left\{ - \int_0^t h(u; x) du \right\} \\ &= \exp \left\{ - \int_0^t h_0(u) \exp(\beta'x) du \right\} \\ &= \exp \left\{ - \exp\{\beta'x\} \int_0^t h_0(u) du \right\}. \end{aligned}$$

例：白血病データ

いま表 15-1 の白血病の例において, $x = x_1$ のみを考える. x_1 はダミー変数で 0 か 1 の値をとる. (37) より次の関係式を得る.

$$\log(-\log S(t; 1)) = \beta_1 + \log(-\log S(t; 0)) \quad (38)$$

すなわち, 比例ハザード性のもとでは, 共変量の値で層別した時の生存関数の 2 重対数 $\log(-\log S)$ は層間で平行になる必要がある. さらに, 生存関数の 2 重対数間の距離はその共変量の効果を表すパラメータの大きさそのものであることも分かる.

時間または時間の対数を横軸にとり, $\log(-\log S)$ をプロットした図形を 2 重対数プロットという. 層間での 2 重対数プロットが平行でないときに, 比例ハザードモデルを適用しては誤った結論を招く恐れがある.

図 15-5 では, 時間の対数を横軸にとり, 表 15-1 の治療群と対象群における生存関数の Kaplan-Meier 推定量に対する 2 重対数プロットを示している. この図から治療効果は時間によってあまり変化しないことが読み取れる. したがって, このデータに対して比例ハザードモデルが適用できそうである. さて, $x = (x_1, x_2)$ とし, 部分尤度 partial likelihood の最大化によってパラメータの値を推定してみると, $\hat{\beta}_1 = 1.294$, $\hat{\beta}_2 = 1.604$ という結果をえる. この $\hat{\beta}_1 = 1.294$ の値が大体図 15-5 における 2 本の曲線間の距離になっていることが確認できる.

カプラン・マイヤー推定量

部分尤度法の解説や、市販のソフトによるパラメータの推定量の求め方などについては専門書（たとえば、大橋・浜田（1995）を参照）に譲ることとする。

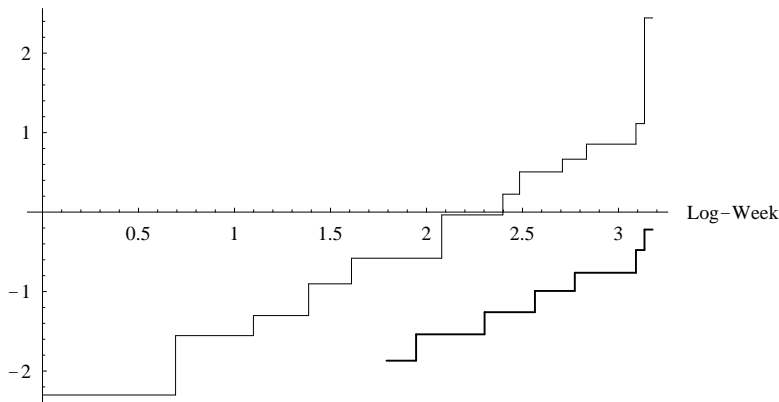


図 15-5: 表 15-1 の治療群（太線）と対象群（細線）におけるカプラン・マイヤー