数理統計

重川 一郎

平成 26 年 8 月 11 日

目 次

第1章　統計データ

1.　データの整理

統計学： 現象の法則性を見出す

記述統計学： ある集団の特徴を記述するために，全体の観測を行い，得られたデータを整理・要約し，そこから現象の法則性を見出す

統計的推測： 一部を観測し，全体の法則性を見出す

度数分布とヒストグラム

例.　試験の成績 $x_1, \ldots, x_{200}$

試験得点の度数分布表

階級	階級値	度数	相対度数	累積度数	累積相対度数
$0 \leq x < 10$	5	6	0.03	6	0.03
$10 \leq x < 20$	15	8	0.04	14	0.07
$20 \leq x < 30$	25	12	0.06	26	0.13
$30 \leq x < 40$	35	21	0.105	47	0.235
$40 \leq x < 50$	45	36	0.18	83	0.415
$50 \leq x < 60$	55	49	0.245	132	0.66
$60 \leq x < 70$	65	30	0.15	162	0.81
$70 \leq x < 80$	75	21	0.105	183	0.915
$80 \leq x < 90$	85	11	0.055	194	0.97
$90 \leq x \leq 100$	95	6	0.03	200	1
合計		200	1.000		

代表値

観測値: $x_1, x_2, \ldots, x_n$

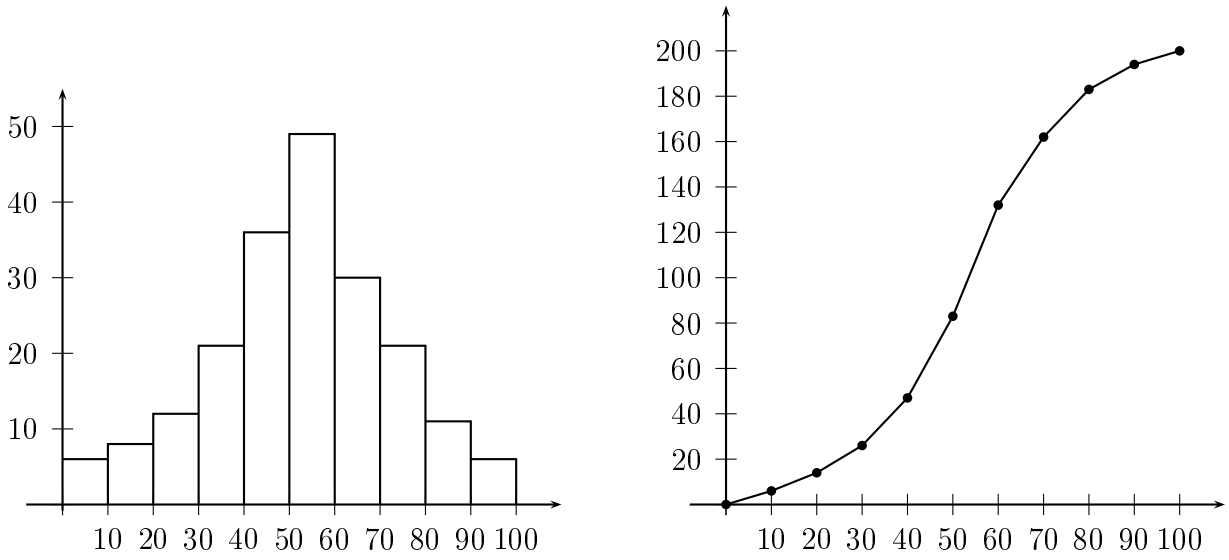


図 1.1: 試験得点のヒストグラムと累積度数グラフ

平均　$\overline{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$

幾何平均　$x_G = \sqrt[n]{x_1 \cdot x_2 \cdot \cdots \cdot x_n}$
地価の上昇、金利など

調和平均　$\dfrac{1}{x_H} = \dfrac{1}{n}\left(\dfrac{1}{x_1} + \dfrac{1}{x_2} + \cdots + \dfrac{1}{x_n}\right)$

メディアン (中央値)　$x_1 \leq x_2 \leq \cdots \leq x_n$ とするとき

$$m = \begin{cases} x_{l+1}, & n \text{ が奇数 } 2l+1 \text{ のとき} \\ \dfrac{x_l + x_{l+1}}{2}, & n \text{ が偶数 } 2l \text{ のとき} \end{cases}$$

モード (最頻値)　一番出現回数の多いもの

ミッド・レンジ　最大と最小の中点　$\frac{1}{2}(\max\{x_i\} + \min\{x_i\})$

調和平均の例
自動車で行きは 40km, 帰りは 50km であったとき，平均時速 v は？
距離を d とすると

$$v = \frac{2d}{\frac{d}{40} + \frac{d}{50}} = \frac{1}{\frac{1}{2}\left(\frac{1}{40} + \frac{1}{50}\right)}$$

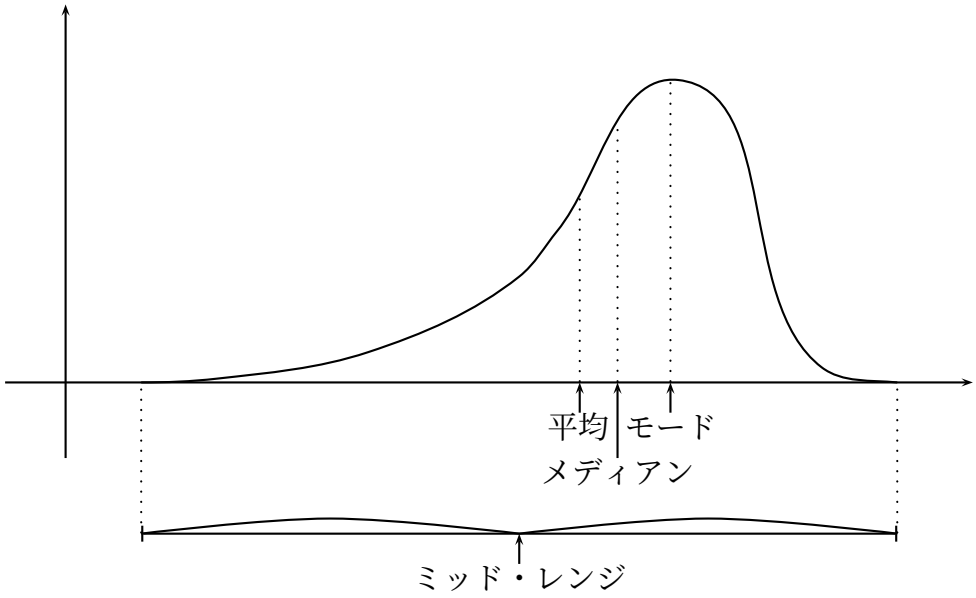


図 1.2: 代表値

$$\frac{1}{v} = \frac{1}{2}\left(\frac{1}{40} + \frac{1}{50}\right) \qquad \therefore\ v = \frac{400}{9} = 44.44\cdots$$

散らばりの尺度

観測値: $x_1, x_2, \ldots, x_n$

レンジ	最大と最小
平均偏差	$\frac{1}{n}(\lvert x_1 - \overline{x}\rvert + \lvert x_2 - \overline{x}\rvert + \cdots + \lvert x_n - \overline{x}\rvert)$
分散	$S^2 = \frac{1}{n}\{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2\}$
標準偏差	$S = \sqrt{S^2}$ (分散の平方根)
標準化	$z_i = \dfrac{x^i - \overline{x}}{S}$, $z_1, z_2, \ldots, z_n$ は 平均 0，分散 1 になる
偏差値	$T_i = 10z_i + 50 = \dfrac{10}{S}(x_i - \overline{x}) + 50$
	平均が 50, 分散が 100， 標準偏差が 10

問題 1.1. 試験をして次のような結果を得た．それぞれについて，平均，分散，標準偏差，偏差値を計算せよ．

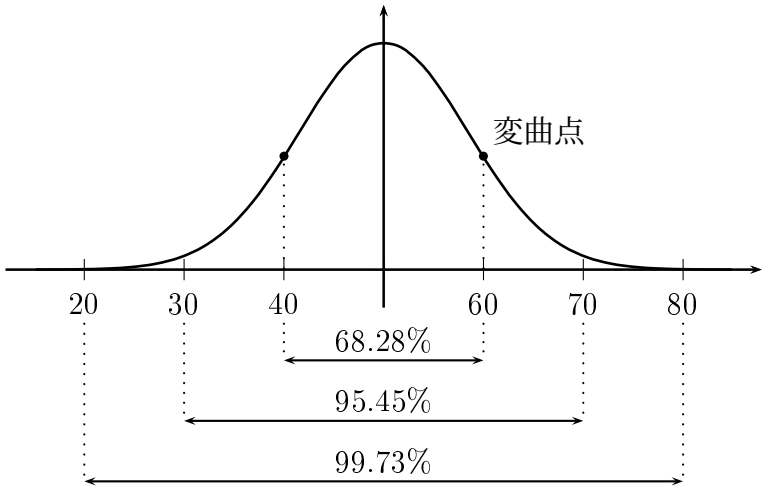


図 1.3: 偏差値 のグラフ

(1)

得点	人数
0	1 人
50	6 人
100	1 人

(2)

得点	人数
0	99 人
100	1 人

解答

(1)

平均　$\dfrac{6 \times 50 + 100}{8} = 50$

分散　$\dfrac{(0-50)^2 + (100-50)^2}{8} = 625$

標準偏差　$\sqrt{625} = 25$

偏差値　0 点の人　$\dfrac{10}{25}(0-50) + 50 = 30$

50 点の人　50

1000 点の人　$\dfrac{10}{25}(100-50) + 50 = 70$

(2)

平均 $\dfrac{100}{100} = 1$

分散 $\dfrac{99 \times (0-1)^2 + (100-1)^2}{100} = \dfrac{99+99^2}{100} = \dfrac{99(1+99)}{100} = 99$

標準偏差 $\sqrt{99} = 9.949\cdots$

偏差値 0 点の人 $\dfrac{10}{9.949}(0-1) + 50 = 48.994\cdots$

1000 点の人 $\dfrac{10}{9.949}(100-1) + 50 = 149.5\cdots$

□

第2章　標本分布

統計的推測 (statistical inference)：母集団から一部を選び出し，それを分析して母集団の推測を行う．

1.　母集団と標本

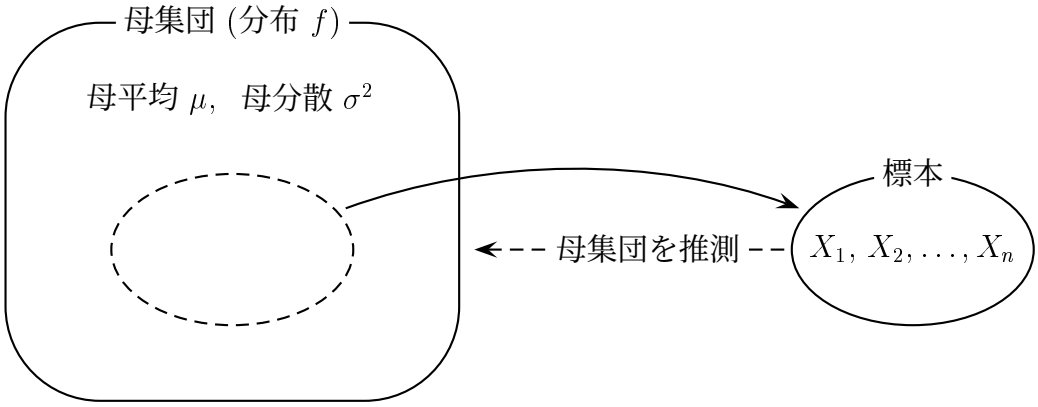


図 2.1: 統計的推測

母集団 (population)：	分析の対象となる集団全体
母集団分布：	母集団の分布
標本 (sample)：	母集団から選び出された要素
母数 (parameter)：	母集団分布を決定するパラメーター

母集団の分布で最も重要なものは正規分布である．このとき，**正規母集団**という．正規分布は平均 μ，分散 σ^2 で特徴付けられる．$N(\mu, \sigma^2)$ とかく．他に，ポアソン分布，二項分布，指数分布などがよく使われる．

数学的に次のように定式化する．

母集団分布	密度関数 (離散のときは確率関数) $f(x)$
母数	母平均 μ, 母分散 σ^2 など
標本	分布 $f(x)$ を持つ**独立**な確率変数 $X_1, X_2, \ldots, X_n$
標本平均 (sample mean)	$\overline{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$
標本分散 (sample variance)	$S^2 = \dfrac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n}$
不偏標本分散 (unbiased sample variance)	$U^2 = \dfrac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1}$

標本平均，標本分散など，標本の関数で，未知の母数を含まないものを**統計量** (statistic) という．

定理 1.1. 分布 f の平均が μ，分散が σ^2 のとき $\overline{X}$ の平均は μ，分散は σ^2/n である．

証明

$$
\begin{aligned}
E[\overline{X}] &= \frac{E[X_1] + E[X_2] + \cdots + E[X_n]}{n} = \frac{n\mu}{n} = \mu \\
V[\overline{X}] &= V\Big(\frac{X_1 + X_2 + \cdots + X_n}{n}\Big) = \frac{1}{n^2} V(X_1 + X_2 + \cdots + X_n) \\
&= \frac{1}{n^2}\{V(X_1) + V(X_2) + \cdots + V(X_n)\} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.
\end{aligned}
$$

□

定理 1.2. 分布 f の平均が μ，分散が σ^2 のとき不偏分散 U^2 の平均は σ^2 である．

証明

$$
E[U^2] = \frac{E[(X_1 - \overline{X})^2] + E[(X_2 - \overline{X})^2] + \cdots + E[(X_n - \overline{X})^2]}{n-1}
$$

である．また

$$
\begin{aligned}
E[(X_1 - \overline{X})^2] &= E[(X_1 - \mu - \overline{X} + \mu)^2] \\
&= E[(X_1 - \mu)^2] - 2E[(X_1 - \mu)(\overline{X} + \mu)] + E[(\overline{X} - \mu)^2] \\
&= \sigma^2 - 2E[(X_1 - \mu)\frac{(X_1 - \mu) + (X_2 - \mu) + \cdots + (X_n - \mu)}{n}] + \frac{\sigma^2}{n} \\
&= \sigma^2 - \frac{2}{n}E[(X_1 - \mu)^2] + \frac{\sigma^2}{n} \\
&= \sigma^2 - \frac{2}{n}\sigma^2 + \frac{\sigma^2}{n}
\end{aligned}
$$

$$= \sigma^2 - \frac{\sigma^2}{n}$$
$$= \frac{n-1}{n}\sigma^2$$

両者から

$$E[U^2] = \frac{nE[(X_1 - \overline{X})^2]}{n-1} = n\frac{n-1}{n}\sigma^2\frac{1}{n-1} = \sigma^2$$

□

統計的推測では $\overline{X}$, U^2 で平均，分散の推定を行う．母数の推定を行う統計量を**推定量** (estimator) という．分散の推定に S^2 ではなく U^2 を用いるのは U^2 の平均が母分散と一致するからである．このように推定量の平均が母数と一致するとき，**不偏推定量**であるという．U^2 を不偏標本分散と呼ぶのはそのためである．

2. 標本分布

以下，推定を行うために必要な標本分布を調べる．まず，正規分布に関する事を纏めておく．$(X_1, X_2, \ldots, X_n)$ を n-次元の（非退化）正規分布とする．すなわち，密度関数が次で与えられる．

$$p(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(V)}} \exp\Big\{\frac{1}{2}(x-m, V^{-1}(x-m))\Big\} \tag{2.1}$$

$m = (m_1, m_2, \ldots, m_n)$ を平均ベクトル，$V = (V_{ij})$ を共分散行列という．次が成り立つ：

$$m_i = E[X_i],$$
$$V_{ij} = Cov(X_i, X_j) = E[(X_i - m_i)(X_j - m_j)].$$

これらの事は特性関数を使うと容易に証明できる．ここで $(X_1, X_2, \ldots, X_n)$ の特性関数 $\varphi(\xi_1, \xi_2, \xi_n)$ は次で定義される：

$$\varphi(\xi_1, \xi_2, \ldots, \xi_n) = E[\exp\{\sum_{j=1}^{n} i\xi_j X_j\}], \quad (\xi_1, \xi_2, \ldots, \xi_n) \in \mathbb{R}^n.$$

特に，密度関数が (2.1) で与えられる正規分布の場合は

$$\varphi(\xi_1, \xi_2, \ldots, \xi_n) = \exp\{i\sum_{j=1}^{n}\xi_j m_j - \frac{1}{2}\sum_{j,k=1}^{n} V_{jk}\xi_j\xi_k\}, \tag{2.2}$$

であることが知られているので，これから平均，共分散が計算できる．(2.2) では，V の正則性は必要ないので，V が非退化の場合も特性関数で正規分布を特徴づけることが出来る．

$X_1, X_2, \ldots, X_n$ が独立で，各 X_j が 1 次元正規分布に従う場合は，密度関数が積になるので，$(X_1, X_2, \ldots, X_n)$ は n-次元正規分布に従う．また $(X_1, X_2, \ldots, X_n)$ が n-次元正規分布

に従えば，$X_1, X_2, \ldots, X_n$ の 1 次結合は，1 次元正規分布に従う．また，(X_1, X_2) が 2 次元正規分布に従い，共分散が 0 であれば

$$\begin{aligned}\varphi(\xi_1,\xi_2) &= \exp\{i\sum_{j=1}^{2}\xi_j m_i - \frac{1}{2}\sum_{j=1}^{2}V_{jj}\xi_j^2\}\\ &= \exp\{i\xi_1 m_1 - \frac{1}{2}V_{11}\xi_1^2\}\exp\{i\xi_2 m_2 - \frac{1}{2}V_{22}\xi_2^2\}\\ &= \varphi_{X_1}(\xi_1)\varphi_{X_2}(\xi_2)\end{aligned}$$

と，特性関数が積になるから独立性が従う．これは正規分布の非常に特殊な性質である．

定理 2.1. X が 1 次元正規分布 $N(\mu,\sigma^2)$ に従うならば，

$$Z = \frac{X-\mu}{\sigma}$$

は標準正規分布 $N(0,1)$ に従う．

また X_1, X_2 が独立で，それぞれ 1 次元正規分布 $N(\mu_1,\sigma_1^2)$, $N(\mu_2,\sigma_2^2)$ に従うとき，その和 X_1+X_2 は 1 次元正規分布, $N(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$ に従う．

さらに $X_1, X_2, \ldots, X_n \sim N(\mu,\sigma^2)$, $\perp\!\!\!\perp$ $\Longrightarrow$ $\overline{X} = \dfrac{X_1+\cdots+X_n}{n} \sim N(\mu,\sigma^2/n)$.

証明　特性関数を計算すればよい．

$$\begin{aligned}E[\exp\{i\xi(X-\mu)/\sigma\}] &= E[\exp\{-i\xi\mu/\sigma\}\exp\{i(\xi/\sigma)X\}]\\ &= \exp\{-i\xi\mu/\sigma\}\exp\{i(\xi/\sigma)\mu + \frac{1}{2}(\xi/\sigma)^2\sigma^2\}\\ &= \exp\{\frac{1}{2}\xi^2\}.\end{aligned}$$

これで Z の分布が $N(0,1)$ であることが示せた．

次に X_1+X_2 の特性関数を計算しよう．独立性から

$$\begin{aligned}E[\exp\{i\xi(X_1+X_2)\}] &= E[\exp\{i\xi X_1\}]E[\exp\{i\xi X_2\}]\\ &= \exp\{i\xi\mu_1 + \frac{\sigma_1^2}{2}\xi^2\}\exp\{i\xi\mu_2 + \frac{\sigma_2^2}{2}\xi^2\}\\ &= \exp\{i\xi(\mu_1+\mu_2) + \frac{\sigma_1^2+\sigma_2^2}{2}\xi^2\}.\end{aligned}$$

これで X_1+X_2 の分布が $N(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$ であることが分かった．

最後に $\overline{X}$ については 定理 1.1 から分布は $N(\mu,\sigma^2/n)$ であることが分かる．　□

ガンマ分布・ベータ分布

定義 2.2. $\alpha > 0$ に対し

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}\,dx \tag{2.3}$$

で定まる関数をガンマ関数，$\alpha, \beta > 0$ に対し

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx \tag{2.4}$$

で定まる関数をベータ関数という．

$\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$ が成り立つことが容易に確かめられる．特に自然数 n に対して $\Gamma(n+1) = n!$ である．

またベータ関数の定義式 (2.4) で $x = \sin^2\theta$ で変数変換すれば $dx = 2\sin\theta\cos\theta$ だから

$$B(\alpha,\beta) = 2\int_0^{\pi/2} \sin^{2(\alpha-1)}\theta\,(1-\sin^2\theta)^{\beta-1}\sin\theta\cos\theta\,d\theta = 2\int_0^{\pi/2} \sin^{2\alpha-1}\theta\,\cos^{2\beta-1}\theta\,d\theta$$

の表示も得られる．これから $B(\frac{1}{2},\frac{1}{2}) = \pi$ が分かる．

定義 2.3. 次の密度関数

$$f_{\alpha,\beta}(x) = \begin{cases} \dfrac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-x/\beta}, & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{2.5}$$

を持つ分布を**ガンマ分布**という．記号で $Ga(\alpha,\beta)$ と表す．

また $[0,1]$ で密度関数

$$\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1} \tag{2.6}$$

を持つ分布を**ベータ分布**という．記号で $Be(\alpha,\beta)$ と表す．

実際に (2.4) が確率密度であることは

$$\begin{aligned}
\int_0^\infty f_{\alpha,\beta}(x)\,dx &= \frac{1}{\Gamma(\alpha)\beta^\alpha}\int_0^\infty x^{\alpha-1}e^{-x/\beta}\,dx \\
&= \frac{1}{\Gamma(\alpha)}\int_0^\infty (x/\beta)^\alpha e^{-x/\beta}\,\frac{dx}{x} \quad (y = x/\beta,\ dy = dx/\beta) \\
&= \frac{1}{\Gamma(\alpha)}\int_0^\infty y^\alpha e^{-y}\,\frac{dy}{y} \\
&= \frac{1}{\Gamma(\alpha)}\int_0^\infty y^{\alpha-1}e^{-y}\,dy \\
&= 1
\end{aligned}$$

から分かる．

命題 2.4.

$$f_{\alpha_1,\beta} * f_{\alpha_2,\beta} = f_{\alpha_1+\alpha_2,\beta}. \tag{2.7}$$

ここで $*$ は合成積

$$f * g(x) = \int_{-\infty}^{\infty} f(x-y)g(y)\,dy$$

を表す.

証明　$x \geq 0$ のとき

$$\begin{aligned}
f_{\alpha_1,\beta} * f_{\alpha_2,\beta}(x) &= \int_{x\geq y, y\geq 0} \frac{1}{\Gamma(\alpha_1)\beta^{\alpha_1}}(x-y)^{\alpha_1-1}e^{-(x-y)/\beta}\frac{1}{\Gamma(\alpha_2)\beta^{\alpha_2}}y^{\alpha_2-1}e^{-y/\beta}\,dy \\
&= \int_0^x \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1+\alpha_2}}(x-y)^{\alpha_1-1}y^{\alpha_2-1}e^{-x/\beta}\,dy \quad (y=tx,\ dy=xdt) \\
&= \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1+\alpha_2}}e^{-x/\beta}\int_0^1 (x-tx)^{\alpha_1-1}(tx)^{\alpha_2-1}x\,dt \\
&= \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1+\alpha_2}}e^{-x/\beta}x^{\alpha_1+\alpha_2-1}\int_0^1 (1-t)^{\alpha_1-1}t^{\alpha_2-1}\,dt \\
&= \frac{B(\alpha_1,\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\frac{1}{\beta^{\alpha_1+\alpha_2}}x^{\alpha_1+\alpha_2-1}e^{-x/\beta} \\
&= \frac{B(\alpha_1,\alpha_2)\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}f_{\alpha_1+\alpha_2,\beta}(x).
\end{aligned}$$

$x < 0$ のとき $f_{\alpha_1,\beta} * f_{\alpha_2,\beta}(x) = 0$ は明らかである.

ここで両辺を積分すると，確率密度であることを用いて

$$\begin{aligned}
\int_{-\infty}^{\infty} f_{\alpha_1,\beta} * f_{\alpha_2,\beta}(x)\,dx &= \frac{B(\alpha_1,\alpha_2)\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\int_{-\infty}^{\infty} f_{\alpha_1+\alpha_2,\beta}(x)\,dx \\
&= \frac{B(\alpha_1,\alpha_2)\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}.
\end{aligned}$$

左辺は

$$\begin{aligned}
\int_{-\infty}^{\infty} f_{\alpha_1,\beta} * f_{\alpha_2,\beta}(x)dx &= \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} f_{\alpha_1,\beta}(x-y)f_{\alpha_2,\beta}(y)\,dy \\
&= \int_{-\infty}^{\infty} f_{\alpha_2,\beta}(y)\,dy \int_{-\infty}^{\infty} f_{\alpha_1,\beta}(x-y)\,dx = 1.
\end{aligned}$$

結局

$$\frac{B(\alpha_1,\alpha_2)\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} = 1$$

も示せ，証明が終わる. □

上の証明中で次の公式

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{2.8}$$

も証明できている．ここで $\alpha=\beta=\frac{1}{2}$ とすれば

$$B\Big(\frac{1}{2},\frac{1}{2}\Big)=\frac{\Gamma(\frac{1}{2})^2}{\Gamma(1)}.$$

これと $B(\frac{1}{2},\frac{1}{2})=\pi$ とを考え合わせれば $\Gamma(\frac{1}{2})=\sqrt{\pi}$ が示せたことになる．

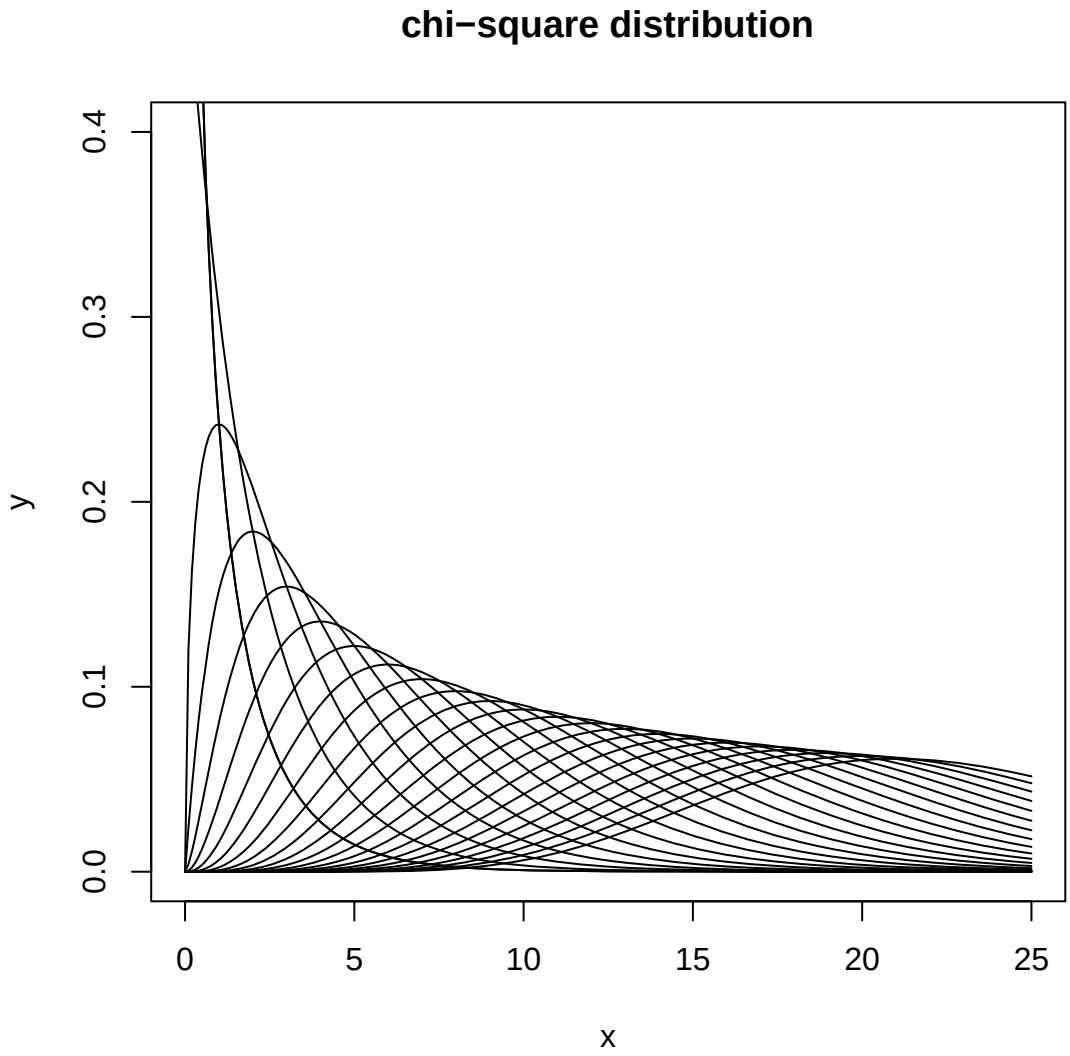
ここで合成積の確率論的な意味を述べておく．X,Y を独立な確率変数で，密度関数 f,g を持つとする．このとき $f*g$ は $X+Y$ の密度関数になっているのである．これを見るには

$$\begin{aligned}
E[F(X+Y)] &= \iint F(x+y)f(x)g(y)\,dx\,dy \\
&\quad u=x+y, v=y \\
&\quad \frac{\partial(x,y)}{\partial(u,v)}=\begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix}=\begin{vmatrix} 1 & 1 \\ 0 & 1 \end{vmatrix}=1 \\
&\quad dx\,dy=\left|\frac{\partial(x,y)}{\partial(u,v)}\right|du\,dv=du\,dv \\
&= \iint F(u)f(u-v)g(v)\,du\,dv \\
&= \int F(u)\,f*g(u)\,du
\end{aligned}$$

に注意すればよい．

カイ 2 乗分布

定義 2.5. n を自然数とするとき，分布 $Ga(n/2,2)$ を**自由度 n の χ^2(カイ 2 乗）分布**という．また記号で $\chi^2(n)$ と表す．



上のカイ 2 乗分布分布の密度のグラフは自由度を 1 から 23 まで動かして描いてある．自由度が 1 のものは原点で発散している．自由度 2 は指数分布であった．自由度が大きくなるとともにグラフは右側に移動しているのを見ることができる．

命題 2.6. $X \sim N(0,1) \Longrightarrow X^2 \sim \chi^2(1)$

証明

$$
\begin{aligned}
E[F(X^2)] &= \int_{-\infty}^{\infty} F(x^2)\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx \\
&= 2\int_{0}^{\infty} F(x^2)\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx \quad (y = x^2,\ dy = 2xdx \ \longrightarrow\ dx = \frac{dy}{2\sqrt{y}}) \\
&= 2\int_{0}^{\infty} F(y)\frac{1}{\sqrt{2\pi}}e^{-y/2}\frac{dy}{2\sqrt{y}} \\
&= \int_{0}^{\infty} F(y)\frac{1}{\sqrt{2\pi}}y^{(1/2)-1}e^{-y/2}\,dy
\end{aligned}
$$

$$= \int_0^\infty F(y)\frac{1}{\Gamma(1/2)2^{1/2}}y^{(1/2)-1}e^{-y/2}\,dy$$
$$= \int_0^\infty F(y)f_{1/2,2}(y)\,dy.$$

□

命題 2.7. $X_1, X_2, \ldots, X_n \sim N(0,1), \perp\!\!\!\perp \implies X_1^2 + X_2^2 + \cdots + X_n^2 \sim \chi^2(n).$
この事実を

$$\chi^2(n) = \underbrace{N(0,1)^2 + N(0,1)^2 + \cdots + N(0,1)^2}_{n}$$

とかく.

証明 X_j^2 の密度関数は $f_{1/2,2}$ で独立であるから,

$$X_1^2 + X_2^2 + \cdots + X_n^2 \sim f_{1/2,2} * f_{1/2,2} * \cdots * f_{1/2,2} = f_{n/2,2}.$$

□

系 2.8. $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2), \perp\!\!\!\perp \implies Y = \dfrac{1}{\sigma^2}\displaystyle\sum_{j=1}^n (X_j - \mu)^2 \sim \chi^2(n).$

証明 X_j^2 の密度関数は $f_{1/2,2}$ で独立であるから,

$$\frac{1}{\sigma^2}(X_j - \mu)^2 = \left(\frac{X_j - \mu}{\sigma}\right)^2, \quad \frac{X_j - \mu}{\sigma} \sim N(0,1)$$

から 命題 2.7 を使えばよい. □

系 2.9. $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2), \perp\!\!\!\perp \implies Y = \dfrac{(\overline{X} - \mu)^2}{\sigma^2/n} \sim \chi^2(1).$

証明 定理 2.1 より $\overline{X} \sim N(\mu, \sigma^2/n)$ であるから 命題 2.6 を使えば明らか. □

命題 2.10. $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2), \perp\!\!\!\perp$
$\implies Y = \dfrac{1}{\sigma^2}\displaystyle\sum_{j=1}^n (X_j - \overline{X})^2 = \frac{n-1}{\sigma^2}U^2 \sim \chi^2(n-1).$
さらに, Y と $\overline{X}$ は独立.

証明 まず $\mu = 0, \sigma^2 = 1$ の場合.

$$Y_j = \sum_{k=1}^n l_{jk}X_k$$

と変換する. $Y_1, Y_2, \ldots, Y_n$ は正規分布に従う. (l_{jk}) をうまく選んで $Y_1, Y_2, \ldots, Y_n$ が次の条件を満たすようにする：

1. $X_1^2 + X_2^2 + \cdots + X_n^2 = Y_1^2 + Y_2^2 + \cdots + Y_n^2$.

2. $Y_n = \sqrt{n}\,\overline{X} = \frac{1}{\sqrt{n}}(X_1 + X_2 + \cdots + X_n)$

3. $Y_1, Y_2, \ldots, Y_n \sim N(0,1)$, $\perp\!\!\!\perp$.

これが成り立つと，

$$Y = \sum_{j=1}^{n}(X_j - \overline{X})^2 = \sum_{j=1}^{n} X_j^2 - n\overline{X}^2 = \sum_{j=1}^{n} Y_j^2 - Y_n^2 = Y_1^2 + Y_2^2 + \cdots + Y_{n-1}^2 \sim \chi^2(n-1)$$

となる．さらに Y と $Y_n = \sqrt{n}\,\overline{X}$ が独立であるから，Y と $\overline{X}$ も独立となって，定理の結論を得る．

従って，上のことを示せば十分である．上で述べていることは 1 次変換 $(X_1, X_2, \ldots, X_n) \mapsto (Y_1, Y_2, \ldots, Y_n)$ が長さ $X_1^2 + X_2^2 + \cdots + X_n^2$ を保ち，$Y_n = \frac{1}{\sqrt{n}}(X_1 + X_2 + \cdots + X_n)$ だから

$$(l_{jk}) = \begin{pmatrix} & \cdots & \cdots & \\ & \cdots & \cdots & \\ \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \end{pmatrix}$$

で，(l_{jk}) が直交行列であればよい．分布密度を

$$\begin{aligned}(X_1, X_2, \ldots, X_n) &\leftrightarrow p(x_1, x_2, \ldots, x_n)\\ (Y_1, Y_2, \ldots, Y_n) &\leftrightarrow q(y_1, y_2, \ldots, y_n)\end{aligned}$$

とすれば

$$p(x_1, x_2, \ldots, x_n) = q(y_1, y_2, \ldots, y_n) \underbrace{\left|\frac{\partial(y_1, y_2, \ldots, y_n)}{\partial(x_1, x_2, \ldots, x_n)}\right|}_{\|\ 1} \quad \text{多変数の変数変換}$$

これから

$$\begin{aligned} p(x_1, x_2, \ldots, x_n) &= \frac{1}{\sqrt{2\pi}^n} e^{-(x_1^2 + x_2^2 + \cdots + x_n^2)/2}\\ q(y_1, y_2, \ldots, y_n) &= \frac{1}{\sqrt{2\pi}^n} e^{-(x_1^2 + x_2^2 + \cdots + x_n^2)/2}\\ &= \frac{1}{\sqrt{2\pi}^n} e^{-(y_1^2 + y_2^2 + \cdots + y_n^2)/2}\\ &= \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-y_j^2/2}. \end{aligned}$$

よって $Y_1, Y_2, \ldots, Y_n$ は独立で $N(0,1)$ に従う.

次に一般の $N(\mu, \sigma^2)$ の場合.

$$Z_j = \frac{X_j - \mu}{\sigma} \sim N(0,1)$$
$$\overline{Z} = \frac{1}{n}(Z_1 + Z_2 + \cdots + Z_n) = \frac{\overline{X} - \mu}{\sigma}$$

であるから

$$\frac{X_j - \overline{X}}{\sigma} = \frac{X_j - \mu}{\sigma} + \frac{\mu - \overline{X}}{\sigma} = Z_j - \overline{Z}$$
$$Y = \frac{1}{\sigma^2}\sum_{j=1}^{n}(X_j - \overline{X})^2 = \sum_{j=1}^{n}(Z_j - \overline{Z})^2.$$

となるので，最初の場合に帰着できる. □

$\overline{X}$ と $X_j - \overline{X}$ が独立であることは，正規分布の性質を使って共分散が 0 であることを示しても分かる．実際

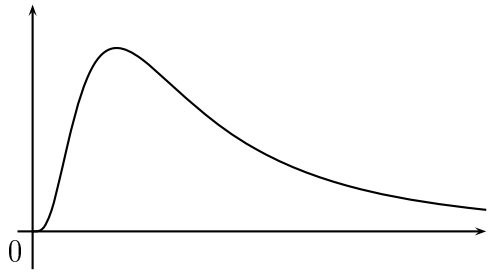
$$\begin{aligned} E[(\overline{X} - \mu)(X_j - \overline{X})] &= E[(\overline{X} - \mu)(X_j - \mu - \overline{X} + \mu)] \\ &= E[(\overline{X} - \mu)(X_j - \mu)] - E[(\overline{X} - \mu)^2] \\ &= \frac{1}{n}E[(X_j - \mu)(X_j - \mu)] - \frac{\sigma^2}{n} = \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0. \end{aligned}$$

F 分布，t 分布

定義 2.11. m, n に対し

$$G_{m,n}(z) = \begin{cases} \dfrac{1}{B(\frac{m}{2}, \frac{n}{2})}\left(\dfrac{m}{n}\right)^{\frac{m}{2}} z^{\frac{m}{2}-1}\left(\dfrac{m}{n}z + 1\right)^{-\frac{m+n}{2}}, & z > 0 \\ 0, & z \leq 0 \end{cases} \tag{2.9}$$

を密度関数に持つ分布を**自由度 (m, n) の F 分布**という．記号で $F(m, n)$ と表す.



F 分布の密度関数のグラフ

定理 2.12. X, Y が独立で，それぞれ自由度 m, n の χ^2 分布をもつとき

$$Z = \frac{X/m}{Y/n} \tag{2.10}$$

は自由度 (m, n) の F 分布をもつ.

補題 2.13. $X \perp\!\!\!\perp Y,\ X \sim p(x),\ Y \sim q(y),\ Y > 0$

$\Longrightarrow \quad Z = \dfrac{aX}{bY}$ の密度関数は

$$r(z) = \int_0^\infty p\Big(\frac{byz}{a}\Big) q(y) \frac{by}{a}\, dy. \tag{2.11}$$

証明

$$\begin{aligned}
E[f(Z)] &= E\Big[f\Big(\frac{aX}{bY}\Big)\Big] \\
&= \int_0^\infty \int_{-\infty}^\infty f\Big(\frac{ax}{by}\Big) p(x)q(y)dxdy \\
&\quad z = \frac{ax}{by},\ u = y \quad \longleftarrow \quad y = u,\ x = \frac{b}{a}yz = \frac{b}{a}uz \\
&\quad \frac{\partial(x,y)}{\partial(z,u)} = \begin{vmatrix} \frac{\partial x}{\partial z} & \frac{\partial x}{\partial u} \\ \frac{\partial y}{\partial z} & \frac{\partial y}{\partial u} \end{vmatrix} = \begin{vmatrix} \frac{b}{a}u & \frac{b}{a}z \\ 0 & 1 \end{vmatrix} = \frac{bu}{a} \\
&\quad dx\, dy = \Big|\frac{\partial(x,y)}{\partial(z,u)}\Big| dz\, du = \frac{bu}{a} dz\, du \\
&= \int_{-\infty}^\infty \int_0^\infty f(z) p\Big(\frac{buz}{a}\Big) q(u) \frac{bu}{a} du\, dz \\
&= \int_{-\infty}^\infty f(z) dz \underbrace{\int_0^\infty p\Big(\frac{buz}{a}\Big) q(u) \frac{bu}{a} du}_{r(z)}.
\end{aligned}$$

□

定理の証明

$$Z = \frac{X/m}{Y/n} = \frac{nX}{mY}$$

$$X \text{ の密度関数は} \quad p(x) = \frac{1}{2^{\frac{m}{2}}\Gamma(\frac{m}{2})} x^{\frac{m}{2}-1} e^{-x/2}$$

$$Y \text{ の密度関数は} \quad q(y) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} y^{\frac{n}{2}-1} e^{-\frac{y}{2}}$$

Z の密度関数は

$$\begin{aligned}
G_{m,n}(z) &= \int_0^\infty p\Big(\frac{m}{n}yz\Big)q(y)\frac{my}{n}dy \\
&= \frac{1}{2^{\frac{m}{2}}2^{\frac{n}{2}}\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}\Big(\frac{m}{n}yz\Big)^{\frac{m}{2}-1}e^{-\frac{1}{2}\frac{m}{n}yz}y^{\frac{n}{2}-1}e^{-\frac{1}{2}y}\frac{m}{n}y\,dy \\
&= \frac{1}{2^{\frac{m}{2}+\frac{m}{2}}\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}\Big(\frac{m}{n}\Big)^{\frac{m}{2}}z^{\frac{m}{2}-1}\underbrace{\int_0^\infty y^{\frac{m}{2}+\frac{n}{2}}e^{-\frac{1}{2}(1+\frac{m}{n}z)y}\frac{dy}{y}}_{*}
\end{aligned}$$

$$\begin{aligned}
&u = \frac{1}{2}(1+\frac{m}{n}z)y \qquad y = 2(1+\frac{m}{n}z)^{-1}u \\
&du = \frac{1}{2}(1+\frac{m}{n}z)dy = \frac{u}{y}dy \quad \therefore \frac{dy}{y} = \frac{du}{u} \\
&= \int_0^\infty 2^{\frac{m+n}{2}}\Big(1+\frac{m}{n}z\Big)^{-\frac{m+n}{2}}u^{\frac{m+n}{2}}e^{-u}\frac{du}{u} \\
&= 2^{\frac{m+n}{2}}\Big(1+\frac{m}{n}z\Big)^{-\frac{m+n}{2}}\Gamma\Big(\frac{m+n}{2}\Big)
\end{aligned}$$

$$\begin{aligned}
\therefore G_{m,n}(z) &= \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}\Big(\frac{m}{n}\Big)^{\frac{m}{2}}z^{\frac{m}{2}-1}\Big(1+\frac{m}{n}z\Big)^{-\frac{m+n}{2}} \\
&= \frac{1}{B(\frac{m}{2},\frac{n}{2})}\Big(\frac{m}{n}\Big)^{\frac{m}{2}}z^{\frac{m}{2}-1}\Big(1+\frac{m}{n}z\Big)^{-\frac{m+n}{2}}.
\end{aligned}$$

よって F 分布が得られた. □

上の事実を

$$F(m,n) = \frac{\chi^2(m)/m}{\chi^2(n)/n}$$

と略記する.

系 2.14. $X_1, X_2, \ldots, X_m$: 同分布の正規分布，$Y_1, Y_2, \ldots, Y_n$: 同分布の正規分布，さらに全て独立で，分散も全て同じ．$\amalg \implies$

$$Z = \frac{\frac{1}{m-1}\{(X_1-\overline{X})^2+\cdots+(X_m-\overline{X})^2\}}{\frac{1}{n-1}\{(Y_1-\overline{Y})^2+\cdots+(Y_n-\overline{Y})^2\}} = \frac{U_X^2}{U_Y^2} \longleftarrow \text{(それぞれの不偏分散)}$$

は自由度 $(m-1, n-1)$ の F 分布をもつ.

ここで，分散は同じであると仮定したが，平均は X_i と Y_j で違ってもよいことを注意しておく.

証明 分散を σ^2 とすると

$$\frac{1}{\sigma^2}\sum(X_i-\overline{X})^2 = \frac{(m-1)U_X^2}{\sigma^2} \sim \chi^2(m-1) \quad (\because \text{命題 2.10})$$

$$\frac{1}{\sigma^2}\sum(Y_j-\overline{Y})^2=\frac{(n-1)U_Y^2}{\sigma^2}\sim\chi^2(n-1)\quad(\because\text{ 命題 2.10})$$
$$\frac{U_Y^2}{U_X^2}\sim\frac{\chi^2(m-1)/(m-1)}{\chi^2(n-1)/(n-1)}=F(m-1,n-1)$$

□

系 2.15. $X_1, X_2, \ldots, X_n$: i.i.d., $\sim N(\mu,\sigma^2)\implies Y=\dfrac{n(\overline{X}-\mu)^2}{U^2}\sim F(1,n-1).$

証明

$$\frac{n(\overline{X}-\mu)^2}{\sigma^2}\sim\chi^2(1)\quad(\because\text{ 系 2.9})$$
$$\frac{(n-1)U^2}{\sigma^2}=\frac{1}{\sigma^2}\sum(X_i-\overline{X})^2\sim\chi^2(n-1)\quad(\because\text{ 命題 2.10})$$

これらは独立であるから

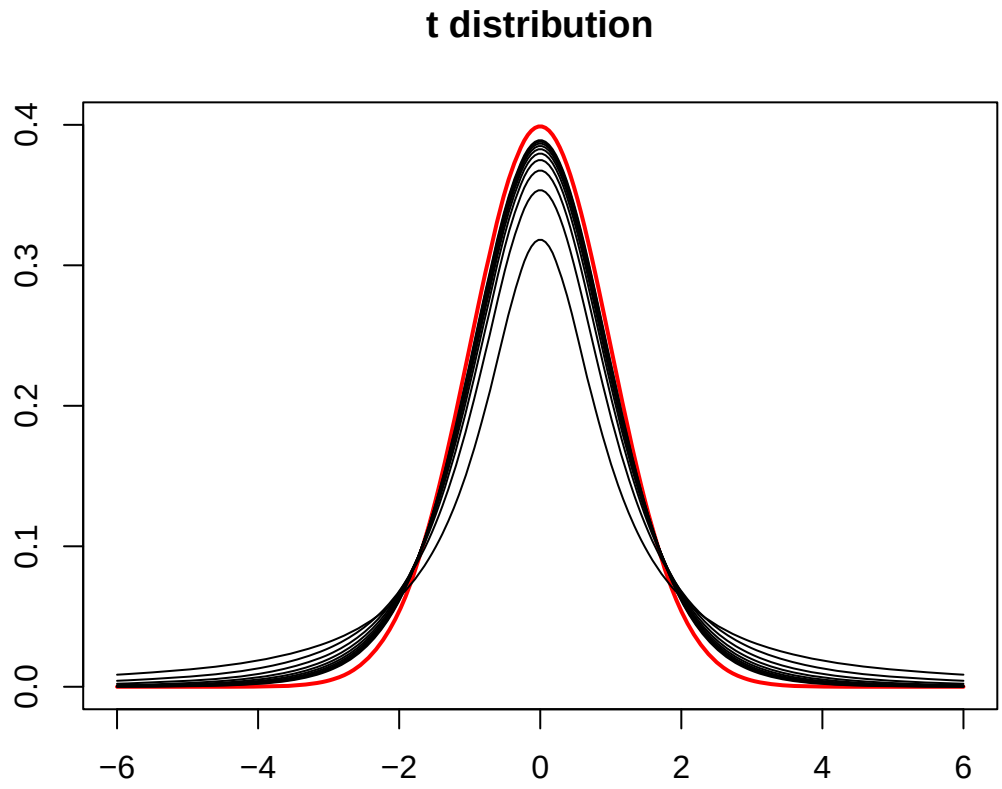
$$\frac{n(\overline{X}-\mu)^2}{U^2}\sim\frac{\chi^2(1)}{\chi^2(n-1)/(n-1)}=F(1,n-1).$$

□

定義 2.16. 次の密度関数を持つ分布を，**自由度 n の t 分布**という.

$$f(x)=\frac{1}{\sqrt{n}B(\frac{1}{2},\frac{n}{2})}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}\tag{2.12}$$

記号で $t(n)$ と表す.



上の t 分布のグラフは，自由度を 1 から 10 まで動かして描いてある．自由度が大きくなるに従って裾野が薄くなり，ピークが高くなる．自由度が大きくなると正規分布に近づく．ピークが最も高いグラフが正規分布のグラフである．

定理 2.17. $X \sim N(0,1),\ Y \sim \chi^2(n),\ X \perp\!\!\!\perp Y \implies \dfrac{X}{\sqrt{\frac{Y}{n}}}$ は自由度 n の t 分布を持つ．

上の事実を

$$t(n) = \frac{N(0,1)}{\sqrt{\chi^2(n)/n}}$$

と表す．

証明

$$E\Big[F\Big(\frac{X}{\sqrt{\frac{Y}{n}}}\Big)\Big] = \int_{-\infty}^{\infty}\int_{0}^{\infty} F\Big(\frac{x}{\sqrt{\frac{y}{n}}}\Big)\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}y^{\frac{n}{2}-1}e^{-\frac{1}{2}y}dy\,dx$$

$$t = \frac{x}{\sqrt{\frac{y}{n}}},\ u = \frac{y}{n} \quad \leftarrow \quad x = t\sqrt{u},\ y = nu$$

$$\frac{\partial(x,y)}{\partial(t,u)} = \begin{vmatrix} \frac{\partial x}{\partial t} & \frac{\partial x}{\partial u} \\ \frac{\partial y}{\partial t} & \frac{\partial y}{\partial u} \end{vmatrix} = \begin{vmatrix} \sqrt{u} & \frac{t}{2\sqrt{u}} \\ 0 & n \end{vmatrix} = n\sqrt{u}$$

$$= \int_{-\infty}^{\infty}\int_{0}^{\infty} F(t)\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2u}{2}}\frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}(nu)^{\frac{n}{2}-1}e^{-\frac{nu}{2}}n\sqrt{u}du\,dt$$

$$= \frac{n^{\frac{n}{2}}}{2^{\frac{n+1}{2}}\sqrt{\pi}\Gamma(\frac{n}{2})}\int_{-\infty}^{\infty} F(t)\underbrace{\left\{\int_{0}^{\infty} e^{-\frac{t^2+n}{2}u}u^{\frac{n-1}{2}}du\right\}}_{*}dt$$

$$v = \frac{n}{2}(1+\frac{t^2}{n})u \qquad dv = \frac{n}{2}(1+\frac{t^2}{n})du$$

$$\int_{0}^{\infty} e^{-v}\Big(\frac{2v}{n(1+\frac{t^2}{n})}\Big)^{-\frac{n-1}{2}}\frac{2}{n(1+\frac{t^2}{n})}dv$$

$$= 2^{\frac{n+1}{2}}n^{-\frac{n+1}{2}}\Big(1+\frac{t^2}{n}\Big)^{-\frac{n+1}{2}}\int_{0}^{\infty} e^{-v}v^{\frac{n+1}{2}-1}dv$$

$$= 2^{\frac{n+1}{2}}n^{-\frac{n+1}{2}}\Big(1+\frac{t^2}{n}\Big)^{-\frac{n+1}{2}}\Gamma\Big(\frac{n+1}{2}\Big)$$

$$= \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n}\sqrt{\pi}\Gamma(\frac{n}{2})}\int_{-\infty}^{\infty} F(t)\Big(1+\frac{t^2}{n}\Big)^{-\frac{n+1}{2}}dt$$

$$= \frac{1}{\sqrt{n}B(\frac{1}{2},\frac{n}{2})}\int_{-\infty}^{\infty} F(t)\Big(1+\frac{t^2}{n}\Big)^{-\frac{n+1}{2}}dt$$

□

命題 2.18. $X_1, X_2, \ldots, X_n$: i.i.d., $\sim N(\mu,\sigma^2) \implies \dfrac{\sqrt{n}(\overline{X}-\mu)}{U} \sim t(n-1)$. ここに U は不変分散 U^2 の平方根である.

証明

$$\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \sim N(0,1), \quad (\because \text{ 定理 2.1})$$
$$\frac{(n-1)U^2}{\sigma^2} \sim \chi^2(n-1), \quad (\because \text{ 命題 2.10})$$

$$\frac{\sqrt{n}(\overline{X}-\mu)}{U} = \frac{\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma}}{\sqrt{\frac{(n-1)U^2}{\sigma^2}\frac{1}{n-1}}} \sim \frac{N(0,1)}{\sqrt{\chi^2(n-1)/(n-1)}} = t(n-1).$$

よって

$$\frac{\sqrt{n}(\overline{X}-\mu)}{U} \sim t(n-1).$$

□

命題 2.19.

$$Z \sim t(n) \implies Z^2 \sim F(1,n).$$

証明 $X \sim N(0,1), Y \sim \chi^2(n), X \perp\!\!\!\perp Y$

$$Z = \frac{X}{\sqrt{\frac{Y}{n}}}, \quad Z^2 = \frac{X^2}{Y/n} \sim \frac{\chi^2(1)}{\chi^2(n)/n} = F(1,n).$$

□

t 分布を使わないで，平方して F 分布を使ってもよい．

一口メモ: No. 1

t-分布はイギリスの統計学者ウィリアム・ゴセット (William Gosset 1876-1937) によって発見された．彼はアイルランドのダブリンにあるビール会社ギネス Guinness で品質管理の仕事に携わっていた．彼はピアソンから統計の教えを受け，品質管理に統計理論を用いていたが，標本が小さいときに分散の推定に S^2 を使うと値がおかしくなることに気が付き，そこから t-分布を発見した．ゴセットはギネスの社員であることから，Student という仮名で論文を発表したので，スチューデントの t-分布とも呼ばれている．

2 項分布と F 分布

2 項分布と F 分布には密接な関係がある．

命題 2.20. 確率変数 X は 2 項分布 $B(N,p)$ に従うとする．このとき $0 \le r \le N$ に対し，$n_1 = 2(r+1)$, $n_2 = 2(N-r)$, $x_0 = \frac{n_2 p}{n_1 q}$ $(q = 1-p)$ として

$$P(X=0) + P(X=1) + \cdots + P(X=r) = \int_{x_0}^{\infty} F(n_1,n_2)(x)\,dx \tag{2.13}$$

が成立する．

また $n_1 = 2(N-r)$, $n_2 = 2(r+1)$, $x_1 = \frac{n_2 q}{n_1 p}$ として

$$P(X=r+1) + P(X=r+2) + \cdots + P(X=N) = \int_{x_1}^{\infty} F(n_1,n_2)(x)\,dx \tag{2.14}$$

が成立する．

証明 まず次の等式を証明する．

$$P(X=0) + P(X=1) + \cdots + P(X=r) = \frac{N!}{r!(N-r-1)!} \int_p^1 y^r (1-y)^{N-r-1}\,dy. \tag{2.15}$$

実際

$$\int_p^n y^r (1-y)^{N-r-1}\,dy = -\left[\frac{1}{N-r} y^r (1-y)^{N-r}\right]_p^1 + \frac{r}{N-r}\int_p^1 y^{r-1}(1-y)^{N-r}\,dy$$

$$= \frac{1}{N-r}p^r(1-p)^{N-r} + \frac{r}{N-r}\int_p^1 y^{r-1}(1-y)^{N-r}\,dy.$$

両辺に $\frac{N!}{r!(N-r-1)!}$ を掛けて

$$\frac{N!}{r!(N-r-1)!}\int_p^n y^r(1-y)^{N-r-1}\,dy$$
$$= \frac{N!}{r!(N-r)!}\frac{1}{N-r}p^r q^{N-r} + \frac{N!}{(r-1)!(N-r)!}\int_p^1 y^{r-1}(1-y)^{N-r}\,dy.$$

これを繰り返せば

$$\frac{N!}{r!(N-r-1)!}\int_p^1 y^r(1-y)^{N-r-1}\,dy$$
$$= \sum_{i=1}^{k}\binom{N}{r-i}p^{r-i}q^{N-r+i} + \frac{N!}{(r-k-1)!(N-r+k)!}\int_p^1 y^{r-k-1}(1-y)^{N-r+k}\,dy$$

が成立する．この式で $k = r-1$, $r-i = j$ とおいて

(2.16)
$$\frac{N!}{r!(N-r-1)!}\int_p^1 y^r(1-y)^{N-r-1}\,dy = \sum_{j=1}^{r}\binom{N}{j}p^j q^{N-j} + \frac{N!}{(N-1)!}\int_p^1 (1-y)^{N-r+k}\,dy$$
$$= \sum_{j=1}^{r}\binom{N}{j}p^j q^{N-j} + (1-p)^N$$
$$= \sum_{j=0}^{r}\binom{N}{j}p^j q^{N-j}$$

ここで $n_1 = 2(r+1)$, $n_2 = 2(N-r)$ すなわち $r = \frac{n_1}{2} - 1$, $N - r = \frac{n_2}{2}$ とすると

$$\frac{N!}{r!(N-r-1)!} = \frac{\Gamma(N+1)}{\Gamma(r+1)\Gamma(N-r)} = \frac{\Gamma(\frac{n_1+n_2}{2})}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})} = \frac{1}{B(\frac{n_1}{2},\frac{n_2}{2})}.$$

上の式は 2 項分布とベータ分布の関係を述べているわけであるが，ベータ分布を変数変換して F 分布が出てくることをみよう．$y = \frac{n_1 x}{n_1 x + n_2}$ で変数変換する：

$$1 - y = \frac{n_1}{n_1 x + n_2}, \quad dy = \frac{n_1 n_2}{(n_1 x + n_2)^2}dx$$

また $y = p$ のとき

$$p = \frac{n_1 x}{n_1 x + n_2},$$
$$pn_1 x + pn_2 = n_1 x$$
$$n_1 x(1-p) = pn_2$$

$$x = \frac{n_2 p}{n_1 p}$$

よって

$$\begin{aligned}\int_p^1 y^r (1-y)^{N-r-1}\, dy &= \int_{x_0}^{\infty} \left(\frac{n_1 x}{n_1 x + n_2}\right)^{\frac{n_1}{2}-1} \left(\frac{n_2}{n_1 x + n_2}\right)^{\frac{n_2}{2}-1} \frac{n_1 n_2}{(n_1 x + n_2)^2}\, dx \\ &= n_1^{n_1/2} n_2^{n_2/2} \int_{x_0}^{\infty} \frac{x^{\frac{n_1}{2}-1}}{(n_1 x + n_2)^{(n_1+n_2)/2}}\, dx\end{aligned}$$

結局 (2.16) の左辺は次に等しい.

$$\frac{n_1^{n_1/2} n_2^{n_2/2}}{B(\frac{n_1}{2}, \frac{n_2}{2})} \int_{x_0}^{\infty} \frac{x^{\frac{n_1}{2}-1}}{(n_1 x + n_2)^{(n_1+n_2)/2}}\, dx$$

これは $F(n_1, n_2)$ 分布の積分である. これで (2.13) が示せた.

(2.14) を示すには, p, と q を入れ替えればよい. すなわち Y を二項分布 $B(N, q)$ に従う確率変数として

$$\begin{aligned}&P(X = r+1) + P(X = r+2) + \cdots + P(X = N) \\ &\qquad = P(Y = 0) + P(Y = 1) + \cdots + P(Y = N - r - 1)\end{aligned}$$

となるので (2.13) の場合に帰着できる. □

第3章　推定

1.　点推定

母集団分布が確率密度 $f(x;\theta)$ を持つとする．$f(x;\theta)$ の形は分かっているが，母数 θ の値が不明であるとする．θ の値を標本 $X_1, X_2,\ldots, X_n$ から推定するのが統計的推定の基本的な問題である．すなわち $X_1, X_2,\ldots, X_n$ の関数として

$$\hat{\theta}_n = g_n(X_1, X_2, \ldots, X_n)$$

で θ を推定する．$\hat{\theta}_n$ は **推定量**と呼ばれている．この推定量は後に出てくる区間推定に対して，**点推定**とよばれる．ここでは，標本の大きさを n を明示的に表すために $\hat{\theta}_n$ と記したが，省略することも多い．また，一般に推定量は $\hat{}$ をつけて表すことにする．

$$E[\hat{\theta}_n] = \theta$$

のとき，$\hat{\theta}_n$ を**不偏推定量**という．推定量のよさを表す重要な指標である．不偏推定量に対してその分散

$$V(\hat{\theta}_n) = E[(\hat{\theta}_n - \theta)^2]$$

は**平均 2 乗予測誤差** (mean square error) と呼ばれるが，これが小さいほうが推定量として望ましい．一般的に分散が最小のものを求めることは難しいが，下からの評価を与えることは出来ることを見ていこう．

フィッシャー情報量

分布族 $\mathcal{F} = \{f(x;\theta)\}_\theta$ が与えられているとき，これらの密度関数を θ の関数とみなすとき**尤度関数** (likelihood function) という．その対数を**対数尤度関数**という：

$$l(\theta) = \log f(x;\theta).$$

$l(\theta)$ は x の関数でもあるが，それを明示するときは $l(\theta;x)$ とかく．l の導関数を計算すると

$$\dot{l}(\theta) = \frac{\dot{f}(x;\theta)}{f(x;\theta)},$$
$$\ddot{l}(\theta) = \frac{\ddot{f}(x;\theta)}{f(x;\theta)} - \left(\frac{\dot{f}(x;\theta)}{f(x;\theta)}\right)^2.$$

となる．θ に関する微分を $\dot{l}$ と，上にドットをつけて表した．

以下 E で 密度関数 $f(x;\theta)$ に関する積分を表すものとする．(あるいは x に $f(x;\theta)$ を密度に持つ確率変数 X を代入して平均をとったと思ってもよい.)

定義 1.1. 分布族 $\{f(x;\theta)\}$ に対して

$$I(\theta) = E[\dot{l}(\theta)^2] = \int_{-\infty}^{\infty} \frac{\dot{f}(x;\theta)^2}{f(x;\theta)}\,dx \tag{1.1}$$

を θ における**フィッシャー情報量** (Fisher's information) という.

次の等式が成立する：

$$E[\dot{l}(\theta)] = 0, \tag{1.2}$$

$$E[\ddot{l}(\theta)] = -I(\theta). \tag{1.3}$$

実際

$$\begin{aligned} E[\dot{l}(\theta)] &= \int_{-\infty}^{\infty} \frac{\dot{f}(x;\theta)}{f(x;\theta)} f(x;\theta)\,dx \\ &= \int_{-\infty}^{\infty} \dot{f}(x;\theta)\,dx \\ &= \frac{d}{d\theta}\int_{-\infty}^{\infty} f(x;\theta)\,dx \\ &= \frac{d}{d\theta}1 \\ &= 0, \end{aligned}$$

$$\begin{aligned} E[\ddot{l}(\theta)] &= E\left[\frac{\ddot{f}(x;\theta)}{f(x;\theta)} - \left(\frac{\dot{f}(x;\theta)}{f(x;\theta)}\right)^2\right] \\ &= \int_{-\infty}^{\infty} \frac{\ddot{f}(x;\theta)}{f(x;\theta)} f(x;\theta)\,dx - E[\dot{l}(\theta)^2] \\ &= \int_{-\infty}^{\infty} \ddot{f}(x;\theta)\,dx - E[\dot{l}(\theta)^2] \\ &= \frac{d^2}{d\theta}\int_{-\infty}^{\infty} f(x;\theta)\,dx - I(\theta) \\ &= -I(\theta). \end{aligned}$$

クラメル-ラオの不等式

さて大きさ n の標本 $X_1, X_2, \ldots, X_n$ をとったとき，その分布密度関数は

$$f_n(x_1, x_2, \ldots, x_n;\theta) = \prod_{j=1}^{n} f(x_j;\theta)$$

で与えられる．これから対数尤度関数 l_n を上と同様に定める

$$l_n(\theta; x_1, x_2, \ldots, x_n) = \log f_n(x_1, x_2, \ldots, x_n;\theta) = \sum_{j=1}^{n} l(x_j;\theta).$$

θ で微分すると

$$\dot{l}_n(\theta; x_1, x_2, \ldots, x_n) = \frac{\dot{f}_n(x_1, x_2, \ldots, x_n; \theta)}{f_n(x_1, x_2, \ldots, x_n; \theta)} = \sum_{j=1}^{n} \dot{l}(\theta; x_j).$$

確率変数 S_n を

$$S_n = \dot{l}_n(\theta; X_1, X_2, \ldots, X_n)$$

で定めると

$$\begin{aligned} E[S_n] &= E[\dot{l}_n(\theta; X_1, X_2, \ldots, X_n)] \\ &= \int_{\mathbb{R}^n} \dot{f}_n(x_1, x_2, \ldots, x_n; \theta) dx_1 \cdots dx_n \\ &= \frac{d}{d\theta} \int_{\mathbb{R}^n} f_n(x_1, x_2, \ldots, x_n; \theta) dx_1 \cdots dx_n \\ &= 0. \end{aligned}$$

また不偏性から

$$\theta = E[\hat{\theta}_n] = \int_{\mathbb{R}^n} g(x_1, x_2, \ldots, x_n) f_n(x_1, x_2, \ldots, x_n; \theta)\, dx_1 \cdots dx_n.$$

両辺を θ で微分して

$$\begin{aligned} 1 &= \int_{\mathbb{R}^n} g(x_1, x_2, \ldots, x_n) \dot{f}_n(x_1, x_2, \ldots, x_n; \theta)\, dx_1 \cdots dx_n \\ &= E[g(X_1, X_2, \ldots, X_n) \dot{l}_n(\theta; X_1, X_2, \ldots, X_n)] \\ &= E[\hat{\theta}_n S_n]. \end{aligned}$$

同時密度関数 $f(x_1, x_2, \ldots, x_n; \theta)$ のフィッシャー情報量は

$$I_n(\theta) = E[\dot{l}_n(\theta)^2] = V(S_n) = V(\sum_{j=1}^{n} \dot{l}(\theta; X_j)) = \sum_{j=1}^{n} V(\dot{l}(\theta; X_j)) = nI(\theta)$$

である.

定理 1.2. (**Cramer-Rao**) 不偏推定量 $\hat{\theta}_n$ の分散に対し次の **Cramer-Rao の不等式**が成立する.

$$V(\hat{\theta}_n) \geq \frac{1}{I_n(\theta)} = \frac{1}{nI(\theta)}. \tag{1.4}$$

等号が成り立つのは, θ に関する微分方程式

$$\dot{l}_n(\theta) = I_n(\theta)(g(x_1, x_2, \ldots, x_n) - \theta) \tag{1.5}$$

が成り立つときである. この式を**有効式** (efficient equation) とよぶ.

証明 $E[S_n] = 0$ であるから

$$Cov(S_n, \hat{\theta}_n) = E[S_n\hat{\theta}_n] - E[S_n]\theta = E[S_n\hat{\theta}_n] = 1.$$

一方 Schwarz の不等式から

$$1 = Cov(S_n, \hat{\theta}_n)^2 \leq V(S_n)V(\hat{\theta}_n) = I_n(\theta)V(\hat{\theta}_n).$$

これで (1.4) が示せた．Schwartz の不等式で等号が成り立つのは S_n, $\hat{\theta}_n$ の間に線形関係 $S_n = a\hat{\theta}_n + b$ が成り立つときである．平均をとると

$$0 = a\theta + b \quad \therefore b = -a\theta.$$

よって

$$S_n = a(\hat{\theta}_n - \theta).$$

Cramer-Rao の不等式で等号が成り立つから

$$1 = Cov(S_n, \hat{\theta}_n) = aV(\hat{\theta}_n) = \frac{a}{I_n(\theta)}.$$

すなわち $a = I_n(\theta)$ である．よって

$$\dot{l}_n(\theta; X_1, \ldots, X_n) = S_n = I_n(\theta)(\hat{\theta}_n - \theta) = I_n(\theta)(g(X_1, \ldots, X_n) - \theta).$$

これから求める結果 (1.5) が得られる． □

有効推定量

定義 1.3. 不等式 (1.4) の右辺 $1/I_n(\theta)$ を Cramer-Rao の下限といい，推定量の分散との比

$$\frac{1}{I_n(\theta)V(\hat{\theta}_n)}$$

を推定量 $\hat{\theta}_n$ の**効率** (efficiency) という．効率は 100 倍して % で表す．

定義 1.4. 推定量 $\hat{\theta}_n$ の効率が 100% のとき，**有効統計量**という．

$\hat{\theta}_n$ が有効統計量であれば，微分方程式 (1.5) を解いて

$$l_n(\theta; x_1, \ldots, x_n) = \int I_n(\theta)\, d\theta\, g(x_1, \ldots, x_n) - \int \theta I_n(\theta)\, d\theta + c(x).$$

すなわち，

$$f_n(x_1, \ldots, x_n; \theta) = \exp\{a(\theta)g(x_1, \ldots, x_n) + b(\theta) + c(x_1, \ldots, x_n)\}. \tag{1.6}$$

ここで，

$$a(\theta) = \int I_n(\theta)\, d\theta, \quad b(\theta) = -\int \theta I_n(\theta)\, d\theta$$

で $c(x_1, \ldots, x_n)$ は θ には関係しない関数である．

定義 1.5. 同時密度関数 f_n が

$$f_n(x_1,\dots,x_n;\theta) = h(g(x_1,\dots,x_n),\theta)\,k(x_1,\dots,x_n) \tag{1.7}$$

と表現されるとき，$\hat{\theta}_n = g(X_1,\dots,X_n)$ を**十分統計量** (sufficient statistics) という．ここで $k(x_1,\dots,x_n)$ が θ に依存しないことが本質的である．(1.6) の関係式から有効統計量は十分統計量である．

十分統計量は θ を推定するのに，十分な情報を持っている，といった意味である．

例 1.1. 正規母集団の有効統計量

母集団が正規分布 $N(\mu,\sigma^2)$ を持ち，分散 σ^2 の場合を考える．密度関数は

$$f(x;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

であるから，対数尤度関数は

$$l(\mu) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}$$

で，微分して

$$\dot{l}(\mu) = (x-\mu)/\sigma^2.$$

よってフィッシャー情報量は

$$I(\mu) = E[\dot{l}(\mu)^2] = E\left[\frac{(X-\mu)^2}{\sigma^4}\right] = \frac{1}{\sigma^2}.$$

n 個の標本 $(X_1, X_2, \dots, X_n)$ の同時密度関数は

$$f_n(x_1,\dots,x_n;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}^n} \exp\left\{-\sum_{j=1}^{n}\frac{(x_j-\mu)^2}{2\sigma^2}\right\}$$

対数をとって

$$l_n(\mu) = -\frac{1}{2}\log(2\pi\sigma^2) - \sum_{j=1}^{n}\frac{(x_j-\mu)^2}{2\sigma^2}.$$

μ で微分して

$$\dot{l}_n(\mu) = \sum_{j=1}^{n}\frac{(x_j-\mu)}{\sigma^2} = nI(\mu)(\overline{x}-\mu).$$

ゆえに有効式 (1.5) が成り立っているので $\overline{X}$ は母平均 μ の有効統計量である．

点推定

母集団の母数を推定することを点推定とよんだ．標本平均 $\overline{X}$, 不偏標本分散 U_n^2 などが代表的なものである．推定量の望ましい性質をまとめると次のようになる．

1. 不偏性: 平均すると真の母数に等しい．
2. 有効性: 推定量が不偏性をみたせば，分散 (平均2乗予測誤差) が小さいほうがよい．分散が最小のものを，有効推定量あるいは最小分散不偏推定量という．
3. 十分性: 母数の情報をすべて含んでいる．
4. 一致性: n を大きくすると真の母数に確率収束する．すなわち任意の $\varepsilon > 0$ に対し
$$P(|\hat{\theta}_n - \theta| > \varepsilon) \to 0 \quad \text{as } n \to \infty.$$
5. 漸近正規性: n が大きいと正規分布に近づく．このとき，元の分布が分からなくても正規分布として扱うことが出来る．

最尤法

推定量を決めるには，いくつか方法があるが，ここでは最尤法について述べる．

例 1.2. 2項分布

成功する確率が p の試行を n 回行って p を推定する問題を考えよう．

$X_1,\ X_2,\ldots,\ X_n$ を成功の確率 p のベルヌーイ列，すなわち i.i.d. で $P(X_j = 1) = p$, $P(X_j = 0) = 1 - p$ となるものとする．結果が $X_1 = r_1,\ X_2 = r_2, \ldots, X_n = r_n$ であったとし，成功の回数を r とする．即ち $r = r_1 + r_2 + \cdots + r_n$．このときの確率を $L(p)$ とおくと

$$L(p) = P(X_1 = r_1, X_2 = r_2, \ldots, X_n = r_n) = p^r(1-p)^{n-r}$$

である．$L(p)$ を**尤度関数**と呼ぶ．これを最大にする p を求め推定量とする．すなわち，起こった事象の確率を最大にする母数 p を標本 $X_1,\ X_2,\ldots,\ X_n$ の関数として表す．$L(p)$ の代わりに，対数尤度 $\log L(p)$ の最大を計算する方が簡単な場合が多い．今の場合

$$\log L(p) = r \log p + (n-r)\log(1-p).$$

対数尤度を p で微分して

$$\frac{d}{dp}\log L(p) = \frac{r}{p} - \frac{n-r}{1-p} = \frac{r(1-p) - (n-r)p}{p(1-p)} = \frac{r - np}{p(1-p)}.$$

これが 0 になるのは $p = \frac{r}{n}$ のときで，$\log L(p)$ は $p = \frac{r}{n}$ のとき最大．

$$\hat{p} = \frac{r}{n} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

を**最尤推定量**と呼ぶ．

例 1.3. 正規分布

母集団の分布が正規分布 $N(\mu, \sigma^2)$ のとき，同時密度関数は

$$f_n(x_1, \ldots, x_n; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}^n} \exp\left\{ -\sum_{j=1}^{n} \frac{(x_j - \mu)^2}{2\sigma^2} \right\}$$

である．尤度関数としては μ, σ^2 の関数とみる．対数をとって対数尤度関数を求めると

$$l_n(\mu, \sigma^2) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \sum_{j=1}^{n} \frac{(x_j - \mu)^2}{2\sigma^2}.$$

最大値を求めるために μ, σ^2 で偏微分して，0 点を探す：

$$\frac{\partial}{\partial \mu} l_n(\mu, \sigma^2) = -\sum_{j=1}^{n} \frac{x_j - \mu}{\sigma^2} = 0,$$

$$\frac{\partial}{\partial \sigma^2} l_n(\mu, \sigma^2) = -\frac{n}{2\sigma^2} + \sum_{j=1}^{n} \frac{(x_i - \mu)^2}{2(\sigma^2)^2} = 0$$

を解いて

$$\mu = \frac{x_1 + \cdots + x_n}{n} = \overline{x},$$

$$\sigma^2 = \frac{1}{n} \sum_{j=1}^{n} (x_j - \mu)^2 = \frac{1}{n} \sum_{j=1}^{n} (x_j - \overline{x})^2.$$

よって (μ, σ^2) の最尤推定量は

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n},$$

$$S^2 = \frac{(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n}.$$

S^2 は不偏性をみたさないから，最尤推定量がつねに不偏性をみたすわけではないことが分かる．

2. 区間推定

真の母数 θ が区間 $[L, U]$ に入る確率が $1 - \alpha$ に等しく（あるいは大きく）なるように，確率変数（統計量）L, U をとる．

$$P(L \leq \theta \leq U) \leq 1 - a.$$

L　　下側信頼限界

U　　上側信頼限界

$1 - \alpha$　**信頼係数**　0.99 や 0.95 などがよく使われる

$[L, U]$　**信頼区間**

正規母集団の母平均の推定（分散が既知の場合）

母分散 σ^2 が分かっているときに母平均を推定する．

信頼係数 $1-\alpha$ から標準正規分布を使って

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

となるように Z_α を定める．Z_α は上側確率 $100\alpha\%$ の**パーセント点**と呼ばれる．このとき

$$L = \overline{X} - \frac{\sigma Z_{\alpha/2}}{\sqrt{n}}$$

$$U = \overline{X} + \frac{\sigma Z_{\alpha/2}}{\sqrt{n}}$$

と取り，$[L, U]$ を信頼係数 $1-\alpha$ の信頼区間という．

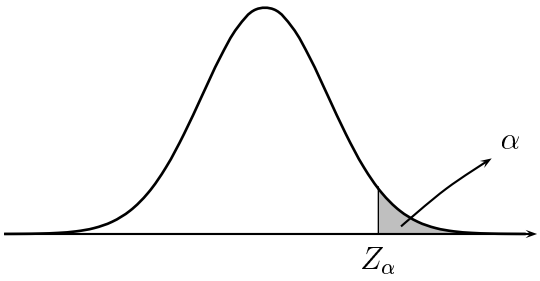


図 3.1: パーセント点

以下，この区間の導出原理を述べる．

$$\overline{X} \sim N(\mu, \sigma^2/n)$$

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

$$P\left(-Z_{\alpha/2} \le \frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \le Z_{\alpha/2}\right) = 1-\alpha$$

$$P\left(\overline{X} - \frac{\sigma Z_{\alpha/2}}{\sqrt{n}} \le \mu \le \overline{X} + \frac{\sigma Z_{\alpha/2}}{\sqrt{n}}\right) = 1-\alpha.$$

パーセント点の例

$$Z_{0.025} = 1.96$$

$$Z_{0.005} = 2.58 \quad (\text{補間すれば } 2.576)$$

例 2.1. 分散 0.25 の正規母集団から，大きさ 25 の標本を取って，標本平均 $\overline{X} = 1.203$ を得た．母平均を信頼係数 95% で推定せよ．

$$\begin{aligned}\overline{X} \pm \frac{\sigma Z_{0.025}}{\sqrt{n}} &= 1.203 \pm \frac{\sqrt{0.25} \times 1.96}{\sqrt{25}} \\ &= 1.203 \pm \frac{1.96}{10} \\ &= 1.203 \pm 0.196\end{aligned}$$

よって信頼区間は $[1.007, 1.399]$.

正規母集団の母分散の推定

$$X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2).$$

σ^2 を推定するのに $U^2 = \frac{1}{n-1}\{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2\}$ を使う.

$$\frac{1}{\sigma^2}\sum_{j=1}^{n}(X_j - \overline{X})^2 \sim \chi^2(n-1)$$

であるから χ^2 分布の上側確率 $100\alpha\%$ のパーセント点を $\chi^2_\alpha(n-1)$ とかくと,

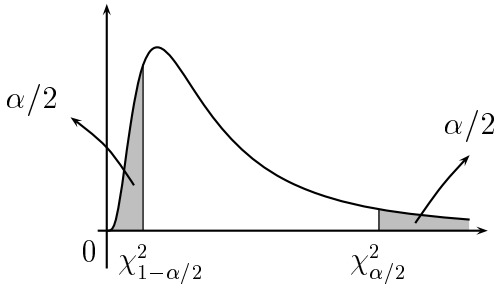


図 3.2: パーセント点

$$P\left(\chi^2_{1-\alpha/2}(n-1) \le \frac{1}{\sigma^2}\sum_{j=1}^{n}(X_j - \overline{X})^2 \le \chi^2_{\alpha/2}(n-1)\right) = 1-\alpha$$

$$P\left(\frac{(n-1)U^2}{\chi^2_{\alpha/2}(n-1)} \le \sigma^2 \le \frac{(n-1)U^2}{\chi^2_{1-\alpha/2}(n-1)}\right) = 1-\alpha.$$

母分散の信頼区間 (信頼係数 $1-\alpha$))

$$\left[\frac{(n-1)U^2}{\chi^2_{\alpha/2}(n-1)}, \frac{(n-1)U^2}{\chi^2_{1-\alpha/2}(n-1)}\right]$$

母平均の推定（分散が未知の場合）

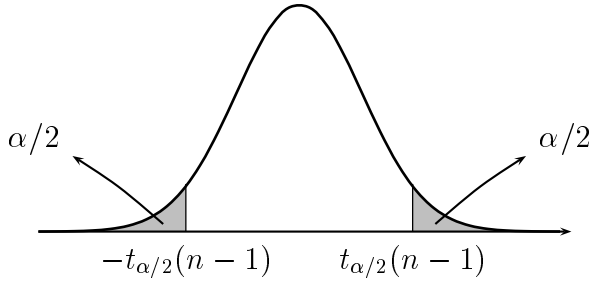
$$X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$$
$$\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$$
$$U^2 = \frac{1}{n-1}\{(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2\}$$

$$\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim N(0,1)$$
$$\frac{(n-1)U^2}{\sigma^2} \sim \chi^2(n-1)$$

$$\frac{\sqrt{n}(\overline{X} - \mu)}{U} = \frac{\dfrac{\sqrt{n}(\overline{X} - \mu)}{\sigma}}{\sqrt{\dfrac{(n-1)U^2}{\sigma^2}\dfrac{1}{n-1}}} \sim \frac{N(0,1)}{\sqrt{\chi^2(n-1)/(n-1)}} = t(n-1)$$

よって

$$\frac{\sqrt{n}(\overline{X} - \mu)}{U} \sim t(n-1)$$



パーセント点

t 分布の上側確率 $100\alpha\%$ のパーセント点を $t_\alpha(n-1)$ とかくと,

$$P\left(-t_{\alpha/2}(n-1) \leq \frac{\sqrt{n}(\overline{X}-\mu)}{U} \leq t_{\alpha/2}(n-1)\right) = 1-\alpha.$$

μ について解いて

$$P\left(\overline{X} - \frac{U}{\sqrt{n}}t_{\alpha/2}(n-1) \leq \mu \leq \overline{X} + \frac{U}{\sqrt{n}}t_{\alpha/2}(n-1)\right) = 1-\alpha.$$

分散が未知の場合の母平均の信頼区間（信頼係数 $1-\alpha$）

$$\left[\overline{X} - \frac{U}{\sqrt{n}}t_{\alpha/2}(n-1),\ \overline{X} + \frac{U}{\sqrt{n}}t_{\alpha/2}(n-1)\right]$$

3.　２標本問題

標本平均の差の推定
2 種類の標本

$$X_1, \ldots, X_m \sim N(\mu_1, \sigma_1^2)$$
$$Y_1, \ldots, Y_n \sim N(\mu_2, \sigma_2^2)$$

$$\overline{X} = \frac{1}{m}(X_1 + \cdots + X_m)$$
$$\overline{Y} = \frac{1}{n}(Y_1 + \cdots + Y_n)$$

$\overline{X} - \overline{Y}$ が母平均の差の推定量.

母平均の差の推定

(1) 母分散が既知

$$\overline{X} \sim N(\mu_1, \frac{\sigma_1^2}{m})$$

$$\overline{Y} \sim N(\mu_2, \frac{\sigma_2^2}{n})$$
$$\overline{X} - \overline{Y} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n})$$

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim N(0,1)$$

$$P\left(-Z_{\alpha/2} \leq \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \leq Z_{\alpha/2}\right) = 1 - \alpha$$

$\mu_1 - \mu_2$ について解いて

$$P\left(\overline{X} - \overline{Y} - \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} Z_{\alpha/2} \leq \mu_1 - \mu_2 \leq \overline{X} - \overline{Y} + \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} Z_{\alpha/2}\right) = 1 - \alpha$$

母分散が既知の場合の母平均の差の信頼区間（信頼係数 $1-\alpha$）

$$\left[\overline{X} - \overline{Y} - \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} Z_{\alpha/2},\ \overline{X} - \overline{Y} + \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} Z_{\alpha/2}\right]$$

(2) 母分散が未知だが等しいとき

$$\overline{X} - \overline{Y} \sim N(\mu_1 - \mu_2, \left(\frac{1}{m} + \frac{1}{n}\right)\sigma^2)$$

$\sigma^2 (= \sigma_1^2 = \sigma_2^2)$ は未知なので推定しなければならない

$$U^2 = \frac{\sum (X_i - \overline{X})^2 + \sum (Y_j - \overline{Y})^2}{m + n - 2}$$
$$= \frac{(m-1)U_X^2 + (n-1)U_Y^2}{m + n - 2}$$

$$\frac{(m+n-2)U^2}{\sigma^2} = \frac{(m-1)U_X^2}{\sigma^2} + \frac{(n-1)U_Y^2}{\sigma^2} \sim \chi^2(m-1) + \chi^2(n-1) = \chi^2(m+n-2)$$

U^2 と $\overline{X} - \overline{Y}$ は 独立である.

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{(\frac{1}{m} + \frac{1}{n})\sigma^2}} \sim N(0,1)$$

$$T = \frac{Z}{\sqrt{\frac{(m+n-2)U^2}{\sigma^2} \frac{1}{m+n-2}}} = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{U\sqrt{(\frac{1}{m} + \frac{1}{n})}} \sim t(m+n-2)$$

母分散が未知だが等しい場合の母平均の差の信頼区間（信頼係数 $1-\alpha$）

$$\left[\overline{X}-\overline{Y}-U\sqrt{\frac{1}{m}+\frac{1}{n}}t_{\alpha/2}(m+n-2),\ \overline{X}-\overline{Y}+U\sqrt{\frac{1}{m}+\frac{1}{n}}t_{\alpha/2}(m+n-2)\right]$$

(3) 母分散が未知で等しくないとき

Welch の近似法がよく使われる

$$T=\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{U_X^2}{m}+\frac{U_Y^2}{n}}}$$

統計量 T を近似的に自由度 ϕ の t 分布とみなす．ϕ は 次の式から定める

$$\frac{(\frac{1}{m}U_X^2+\frac{1}{n}U_Y^2)^2}{\phi}=\frac{(\frac{1}{m}U_X^2)^2}{m-1}+\frac{(\frac{1}{n}U_Y^2)^2}{n-1}$$

母分散が未知で等しくない場合の母平均の差の信頼区間（信頼係数 $1-\alpha$）

$$\left[\overline{X}-\overline{Y}-\sqrt{\frac{U_X^2}{m}+\frac{U_Y^2}{n}}t_{\alpha/2}(\phi),\ \overline{X}-\overline{Y}+\sqrt{\frac{U_X^2}{m}+\frac{U_Y^2}{n}}t_{\alpha/2}(\phi)\right]$$

母分散の比の推定

$$X_1,\ldots,X_m\sim N(\mu_1,\sigma_1^2)$$
$$Y_1,\ldots,Y_n\sim N(\mu_2,\sigma_2^2)$$

$$\overline{X}=\frac{1}{m}(X_1+\cdots+X_m)$$
$$\overline{Y}=\frac{1}{n}(Y_1+\cdots+Y_n)$$
$$U_X^2=\frac{1}{m-1}\{(X_1-\overline{X})^2+\cdots+(X_m-\overline{X})^2\}$$
$$U_Y^2=\frac{1}{n-1}\{(Y_1-\overline{Y})^2+\cdots+(Y_m-\overline{Y})^2\}$$

$$\frac{(m-1)U_X^2}{\sigma_1^2}\sim\chi^2(m-1)$$
$$\frac{(n-1)U_Y^2}{\sigma_2^2}\sim\chi^2(n-1)$$

$$\frac{U_X^2/\sigma_1^2}{U_Y^2/\sigma_2^2} \sim \frac{\chi^2(m-1)/(m-1)}{\chi^2(n-1)/(n-1)} \sim F(m-1, n-1)$$

信頼係数を $1-\alpha$ として

$$P\left(F_{1-\alpha/2}(m-1,n-1) \leq \frac{U_X^2}{U_Y^2}\frac{\sigma_2^2}{\sigma_1^2} \leq F_{\alpha/2}(m-1,n-1)\right) = \alpha$$
$$P\left(\frac{U_Y^2}{U_X^2}F_{1-\alpha/2}(m-1,n-1) \leq \frac{\sigma_2^2}{\sigma_1^2} \leq \frac{U_Y^2}{U_X^2}F_{\alpha/2}(m-1,n-1)\right) = \alpha.$$

よって

母分散の比 σ_2^2/σ_1^2 の信頼区間 (信頼係数 $1-\alpha$)

$$\left[\frac{U_Y^2}{U_X^2}F_{1-\alpha/2}(m-1,n-1),\ \frac{U_Y^2}{U_X^2}F_{\alpha/2}(m-1,n-1)\right]$$

ここで F 分布のパーセント点に関して,

$$F(m,n) \sim \frac{1}{F(n,m)}$$
$$F_{1-\alpha}(m,n) = \frac{1}{F_\alpha(n,m)}$$

であるから，$F_{1-\alpha}(m,n)$ は逆数で計算する.

問題

1. 平均 μ の指数分布は密度関数

$$f(x) = \begin{cases} \frac{1}{\mu}e^{-x/\mu}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

をもつ．このとき平均 μ の，点推定，区間推定の理論を正規分布の場合と同じように作れ．ただし，点推定は，最尤法によるものとする.

2. パラメーター λ のポアソン分布は次の確率分布を持つ.

$$P(X=k) = \frac{\lambda^k}{k!}e^{-\lambda}, \quad k = 0, 1, 2, \ldots.$$

λ を推定するのに，推定量として，平均 $\overline{X} = \dfrac{X_1 + \cdots + X_n}{n}$ を用いたとき，これは不偏性をみたすことを示せ．また効率を求めよ.

3. 次のデータはある小学校の生徒を無作為に 10 人選んだときの身長である．母集団は正規分布であることを仮定して，身長の平均を信頼係数 95% で区間推定せよ.

身長 (cm)	137	120	131	124	127	140	123	116	124	108

第4章　仮説検定

1.　検定の考え方

この節では仮説検定を扱うが，それがどのように行われるかまず例を見ていくことにする．

2項分布の検定

例 1.1. サイコロを180回投げて1の目が39回出た．このサイコロは正常か？

この問題は，1の出る確率が $\frac{1}{6}$ であるかどうか調べよといっている訳である．1の出る確率を p として仮説「$p=\frac{1}{6}$ である」を立てる．数学的には

H_0: $p=\frac{1}{6}$ である　　$\cdots$ 帰無仮説

H_1: $p=\frac{1}{6}$ ではない　$\cdots$ 対立仮説

$p=\frac{1}{6}$ として計算を進める．X を180回投げたときの1の出た回数とする．$X\sim Bin(180,\frac{1}{6})$ である．

$$E[X]=30,\quad V(X)=180\times p\times(1-p)=180\times\frac{1}{6}\times\frac{5}{6}=25=5^2.$$

30が平均であるので，

$|X-30|\le c$　$\cdots$ H_0 は正しいと判断 ＝ 採択

$|X-30|>c$　$\cdots$ H_0 は誤りと判断 ＝ 棄却

c は，有意水準 α を与えて

$$P(|X-30|>c)=\alpha$$

となるように定める．$\alpha=0.05,0.01$ などがよくとられる．$W=\{x;\,|x-30|>c\}$ を棄却域という．c を求めるのに，X は近似的に $N(30,5^2)$ に従うから，これで計算をする．$Z\sim N(0,1)$ として

$$P(|X-30|>c)=P\left(\left|\frac{X-30}{5}\right|>\frac{c}{5}\right)\fallingdotseq P\left(|Z|>\frac{c}{5}\right)=\alpha.$$

今 $\alpha=0.05$ とすると $P(|Z|>1,96)=0.05$ であるから

$$\frac{c}{5}=1.96\quad\therefore c=1.96\times5=9.8\quad\therefore P(|X-30|>9.8)=0.05.$$

よって

$$X<30-9.8=20.2\quad\text{or}\quad X>30+9.8=39.8$$

ならば棄却する．

注意 1.1. ここでは 1 の目だけに着目したが，より正確な検定を行うためには，全ての目の出方を調べた方がよい．そのときには後に第 3 節で述べる適合度検定を使う．

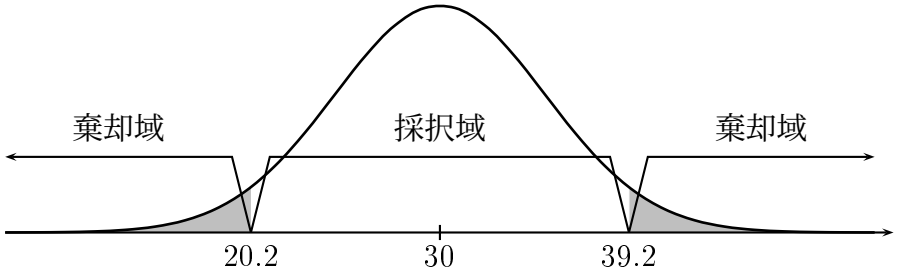


図 4.1: 棄却域と採択域

今の場合は 39 だから採択（棄却できない）.

注意 1.2. 棄却: 帰無仮設 H_0 が正しいとして計算して，非常に確率の低いことが起こった場合に棄却する．例外的なことが起こったので仮説が間違っていると判断するわけである．

採択: 積極的に支持しているわけではない．棄却するには根拠が弱い.（疑わしいが証拠が不十分という状況.）仮説 H_0 が採択されるときは何もいえないというべきで，それゆえ帰無仮説とよばれる．棄却できるときに初めて積極的な意味を持つ．(気持ちとしては，対立仮説 H_1 の方を言いたい．)

検定の手順

検定の手順

1. 仮説の設定 (帰無仮説と対立仮説)
2. 検定統計量の選定
3. 有意水準を定める $(\alpha = 0.05, 0.01)$
4. 棄却域の計算
5. 採択・棄却の決定

検定の種類

両側検定

$$\left.\begin{array}{l} H_0 : \ \theta = \theta_0 \\ H_1 : \ \theta \neq \theta_0 \end{array}\right\} \quad \longrightarrow \quad \theta = \theta_0 \text{ として } P(|\hat{\theta} - \theta_0| > c) = \alpha \text{ となるよう } c \text{ を定める}$$

片側検定

$$\left.\begin{array}{l} H_0: \ \theta = \theta_0 \\ H_1: \ \theta > \theta_0 \end{array}\right\} \longrightarrow \quad \theta = \theta_0 \text{ として } P(\hat{\theta} - \theta_0 > c) = \alpha \text{ となるよう } c \text{ を定める}$$

$$\left.\begin{array}{l} H_0: \ \theta = \theta_0 \\ H_1: \ \theta < \theta_0 \end{array}\right\} \longrightarrow \quad \theta = \theta_0 \text{ として } P(\hat{\theta} - \theta_0 < c) = \alpha \text{ となるよう } c \text{ を定める}$$

それぞれ $\{x;\, |x - \theta_0| > c\}$, $\{x;\, x - \theta_0 > c\}$, $\{x;\, x - \theta_0 < c\}$ が棄却域.

誤りの種類

	H_0 が正しい	H_0 が誤り
採択		第二種の誤り
棄却	第一種の誤り (有意水準 α)	(検出力)

第一種の誤りと第二種の誤りの確率を同時に小さくする事は出来ない．第一種の誤りの確率（有意水準）を固定し，第二種の誤りの確率を出来るだけ小さくするように棄却域を設定する.

2. 正規母集団の検定

以下，母集団は正規分布に従うとする

平均の検定

$X_1, X_2, \ldots, X_n$: 標本

$$H_0: \ \mu = \mu_0$$
$$H_1: \ \mu \neq \mu_0$$

[A] 分散 σ^2 が既知の場合

$$Z = \frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma} \sim N(0, 1).$$

棄却域: $W = \{z;\, |z| > Z_{\alpha/2}\}$. したがって

$$Z \in W \Rightarrow \text{ 棄却}$$
$$Z \notin W \Rightarrow \text{ 採択}$$

片側検定の場合は

$$\left.\begin{array}{ll} H_0: & \mu = \mu_0 \\ H_1: & \mu > \mu_0 \end{array}\right\} \quad \cdots \quad W = \{z;\, z > Z_\alpha\}$$

$$\left.\begin{array}{ll} H_0: & \mu = \mu_0 \\ H_1: & \mu < \mu_0 \end{array}\right\} \quad \cdots \quad W = \{z;\, z < -Z_\alpha\}.$$

[B] 分散 σ^2 が未知の場合

この場合は分散の推定値として $U^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2$ を使う.

$$T = \frac{\sqrt{n}(\overline{X} - \mu_0)}{U} \sim t(n-1),$$
$$W = \{t;\, |t| > t_{\alpha/2}(n-1)\} \ \cdots \text{棄却域}$$

問題 2.1. A 社は，新しく開発した電球が 1700 時間より長い寿命があると主張している．10 個を選んで次のデータを得た．A 社の主張は正しいと言ってよいか．有意水準 5%で検定せよ．

1681, 1580, 1661, 1743, 1715, 1576, 1653, 1805, 1982, 1783

解答

$$H_0 : \mu = 1700$$
$$H_1 : \mu > 1700$$

と仮説を立てる．データから

$$\overline{X} = 1717.9, \quad U = 120.2659, \quad t_{0.05}(9) = 1.833$$
$$T = \frac{\sqrt{10}(1717.9 - 1700)}{120} = 0.470\cdots$$
$$W = \{t;\, t > t_{0.05}(9)\} = \{t > 1.833\}.$$

$0.470 \leq 1.833$ だから仮設は棄却できない．すなわち採択.

この場合 $H_0 : \mu = 1700$ が採択されたということは，寿命が 1700 時間であるということであり，それより長いと主張している A 社の主張は結論的には認められないわけである．しかしながら帰無仮設が採択される場合は積極的な意味ではないから「なんとも言えない」というのが関の山である． □

分散の検定

$X_1, X_2, \ldots, X_n$: 標本

$$\left\{\begin{array}{ll} H_0: & \sigma^2 = \sigma_0^2 \\ H_1: & \sigma^2 \neq \sigma_0^2 \end{array}\right.$$

$$Y = \frac{(n-1)U^2}{\sigma_0^2} \sim \chi^2(n-1)$$
$$W = \{y;\, y < \chi^2_{1-\alpha/2}(n-1), \quad y > \chi^2_{\alpha/2}(n-1)\}$$
$$Y \in W \ \cdots \ \text{棄却}$$
$$Y \notin W \ \cdots \ \text{採択}$$

等平均の検定

$X_1, X_2, \ldots, X_m \sim N(\mu_1, \sigma_1^2), \quad Y_1, Y_2, \ldots, Y_n \sim N(\mu_2, \sigma_2^2).$

$$\begin{cases} H_0 : \ \mu_1 = \mu_2 \\ H_1 : \ \mu_1 \neq \mu_2 \end{cases}$$

[A] 分散 σ_1^2, σ_2^2 が既知の場合

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim N(0,1)$$
$$W = \{z;\, |z| > Z_{\alpha/2}\}$$

[B] 分散が未知で $\sigma_1^2 = \sigma_2^2$ の場合

$$U^2 = \frac{1}{m+n-2}\{\sum_{j=1}^{m}(X_j - \overline{X})^2 + \sum_{j=1}^{n}(Y_j - \overline{Y})^2\}$$
$$T = \frac{\overline{X} - \overline{Y}}{U\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t(m+n-2)$$
$$W = \{t;\, |t| > t_{\alpha/2}(m+n-2)\}$$

[C] 分散 σ_1^2, σ_2^2 が未知の場合
[A] の場合の σ_1^2, σ_2^2 を

$$U_X^2 = \frac{1}{m-1}\sum_{j=1}^{m}(X_j - \overline{X})^2$$
$$U_Y^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \overline{Y})^2$$

で推定し，検定統計量

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{U_X^2}{m} + \frac{U_Y^2}{n}}}$$

を用いる．この統計量 T を近似的に自由度 ϕ の t 分布とみなす．ϕ は 次の式から定める

$$\frac{(\frac{1}{m}U_X^2 + \frac{1}{n}U_Y^2)^2}{\phi} = \frac{(\frac{1}{m}U_X^2)^2}{m-1} + \frac{(\frac{1}{n}U_Y^2)^2}{n-1}.$$

この方法を Welch's test とよぶ.

等分散の検定

$X_1, X_2, \ldots, X_m \sim N(\mu_1, \sigma_1^2)$,　$Y_1, Y_2, \ldots, Y_n \sim N(\mu_2, \sigma_2^2)$.

$$\begin{cases} H_0 : \ \sigma_1^2 = \sigma_2^2 \\ H_1 : \ \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

H_0 の下で

$$F = \frac{U_X^2}{U_Y^2} \sim F(m-1, n-1)$$

$$W = \{f;\, f < \frac{1}{F_{\alpha/2}(n-1, m-1)}, \quad f > F_{\alpha/2}(m-1, n-1)\}$$

3. χ^2 検定

χ^2 検定と呼ばれる検定をいくつか論じる.

適合度検定

分布の型を検定する

例： 正規分布に従っているか?

ポアソン分布に従っているか?

[A] 未知母数を含まない場合

事象	A_1	A_2	$\cdots$	A_k	計
生起確率	p_1	p_2	$\cdots$	p_k	1
期待度数	np_1	np_2	$\cdots$	np_k	n
観測度数	n_1	n_2	$\cdots$	n_k	n

帰無仮説 H_0:　$P(A_i) = p_i$

対立仮説 H_1:　$\exists i \ \ P(A_i) \neq p_i$

$$X = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \qquad \text{自由度 } k-1 \text{ の } \chi^2 \text{ 分布で近似できる}$$

$$W = \{x;\, x > \chi_\alpha^2(k-1)\}$$

この検定を **χ^2 適合度検定**という

[B] 未知母数を含む場合

F_θ: 未知母数 $\theta = (\theta_1, \theta_2, \ldots, \theta_t)$ を含む分布

θ を $\hat{\theta}$ で推定（最尤推定量）

事象 A_i の確率は $P(A_i) = p_i = p_i(\theta)$ であるが，θ が未知であるので，生起確率の推定量として $\hat{\hat{p}}_i = p_i(\hat{\theta})$ を用いる．後は未知母数を含まない場合と同じで，

帰無仮説 H_0: 母集団の分布は F_θ である

対立仮説 H_1: 母集団の分布は F_θ ではない

$$X = \sum_{i=1}^{k} \frac{(n_i - n\hat{\hat{p}}_i)^2}{n\hat{\hat{p}}_i} \qquad \text{自由度 } k-t-1 \text{ の } \chi^2 \text{ 分布で近似できる}$$

$$W = \{x;\, x > \chi^2_\alpha(k-t-1)\}$$

正規分布の場合であれば平均 μ, 分散 σ^2 の母数を持ち，A_i は適当にクラスを分け，$A_i = \{a_i < Z \le a_{i+1}\}$ などとする．

独立性の検定

喫煙と性別，あるいは政党の支持者と年齢といった二つの要因が独立であるかどうかを考えよう．すなわち 2 つの要因 A と B があって，これらが独立であるかどうかを検定することを考える．

A 要因は r 水準，B 要因は c 水準に分かれているとしよう．A 要因と B 要因の確率分布を次で与える．

A \ B	B_1	B_2	$\cdots$	B_c	行周辺分布
A_1	p_{11}	p_{12}	$\cdots$	p_{1c}	$p_{1\cdot}$
A_2	p_{21}	p_{22}	$\cdots$	p_{2c}	$p_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_r	p_{r1}	p_{r2}	$\cdots$	p_{rc}	$p_{r\cdot}$
列周辺分布	$p_{\cdot 1}$	$p_{\cdot 2}$	$\cdots$	$p_{\cdot c}$	1

行周辺分布: $p_{i\cdot} = \sum_{j=1}^{c} p_{ij}$

列周辺分布: $p_{\cdot j} = \sum_{i=1}^{r} p_{ij}$

全確率: $1 = \sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij}$

確率の意味は

$$p_{ij} = P(A_i \cap B_j), \quad p_{i\cdot} = P(A_i), \quad p_{\cdot j} = P(B_j)$$

要因 A, B が独立であるという仮説は

帰無仮説 H_0: $p_{ij} = p_{i\cdot}p_{\cdot j} \quad i = 1, \ldots, r;\ j = 1, \ldots, c$

対立仮説 H_1: それ以外

観測データとして次のものが得られたとする．

A \ B	B_1	B_2	$\cdots$	B_c	行周辺度数
A_1	n_{11}	n_{12}	$\cdots$	n_{1c}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2c}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	$n_{r\cdot}$
列周辺度数	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot c}$	n

行周辺度数: $n_{i\cdot} = \sum_{j=1}^{c} n_{ij}$

列周辺度数: $n_{\cdot j} = \sum_{i=1}^{r} n_{ij}$

全度数: $n = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}$

それぞれ推定量としては

$$\hat{p}_{ij} = \frac{n_{ij}}{n}, \quad \hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n}, \quad \hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n}$$

であるから，統計量としては

$$X = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - \frac{n_{i\cdot}n_{\cdot j}}{n})^2}{\frac{n_{i\cdot}n_{\cdot j}}{n}}$$

を取る．n が十分大きいときには近似的に $\chi^2((r-1)(c-1))$ 分布に従う．自由度がこうなるのは $rc-1$ から，推定した $c-1, r-1$ だけ少ないからである．

以上のことから棄却域は

$$W = \{x;\, x > \chi^2_\alpha((r-1)(c-1))\}$$

で定めればよい．

第5章　統計解析

1.　回帰分析

線型回帰分析

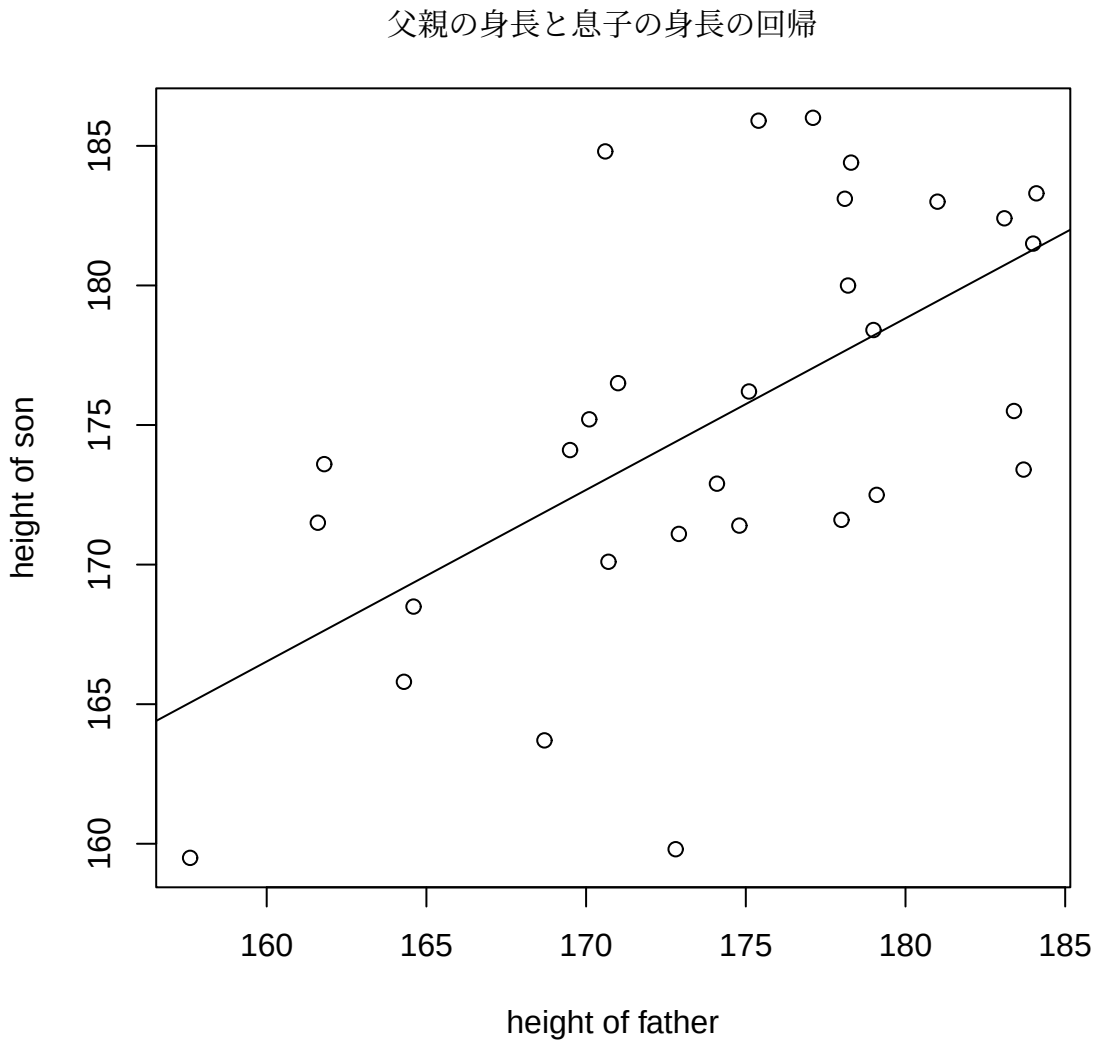
親子の身長，年齢と血圧，高校の成績と大学の成績などの間には前者が，後者に影響を与えるという関係が考えられる．このように一つの変数と他の変数との間の関係を調べることを回帰分析という．とくにここでは線型関係にに限ることにする．この場合を**線形回帰分析**という．

具体例を見ながら考えてみよう．次はピアソンが用いた父親の身長と，息子の身長の30人のデータである．

父親の身長と息子の身長

	1	2	3	4	5	6	7	8	9	10
父親 (cm)	183.7	178.0	178.3	178.1	171.0	157.6	179.1	179.0	170.7	164.6
息子 (cm)	173.4	171.6	184.4	183.1	176.5	159.5	172.5	178.4	170.1	168.5
	11	12	13	14	15	16	17	18	19	20
父親 (cm)	172.9	172.8	184.1	169.5	177.1	178.2	174.1	170.1	170.6	183.1
息子 (cm)	171.1	159.8	183.3	174.1	186.0	180.0	172.9	175.2	184.8	182.4
	21	22	23	24	25	26	27	28	29	30
父親 (cm)	184.0	175.1	168.7	161.6	183.4	174.8	175.4	164.3	181.0	161.8
息子 (cm)	181.5	176.2	163.7	171.5	175.5	171.4	185.9	165.8	183.0	173.6

このデータの散布図と，回帰直線を図示すると，以下のようになる．ここに書き加えられているのが回帰直線で，この求め方を述べていく．



$(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$ を得られたデータとし，x の値が Y に影響を与えているとする．すなわち次の関係があるものとする:

$$Y_j = a + bx_j + \varepsilon_j.$$

x_j は確率変数ではなく確定した値として扱い，**説明変数**（独立変数）という．それに対して，Y を**被説明変数**（従属変数）という．ε_j は**誤差項**（擾乱項） error と呼ばれる．ε_j は独立，同分布を仮定する．以下では正規分布 $N(0, \sigma^2)$ に従うものと仮定する．

$$y = a + bx$$

を**回帰曲線**とよび，a, b を**回帰係数**という．a, b を推定（あるいは検定）することがここでの主題である．

条件から

$$Y_j \sim N(a + bx_j, \sigma^2).$$

尤度関数 i.e., $(Y_1, Y_2, \ldots, Y_n)$ の同時密度関数は

$$L(a,b) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y_j - a - bx_j)^2}{2\sigma^2} \right\}$$

であり，対数尤度関数は

$$l(a,b) = \log L(a,b) = -n \log \sqrt{2\pi\sigma^2} - \sum_{j=1}^{n} \frac{(y_j - a - bx_j)^2}{2\sigma^2}.$$

これを最大化するには $\sum_j (y_j - a - bx_j)^2$ を最小化すればよい．結局最小 2 乗法で求めることになる．計算のために記号を準備しておこう．それぞれの分散，共分散の記号を決めておく．

$$S_x^2 = \frac{1}{n} \sum_{j=1}^{n} (x_j - \overline{x})^2,$$

$$S_y^2 = \frac{1}{n} \sum_{j=1}^{n} (y_j - \overline{y})^2,$$

$$S_{xy} = \frac{1}{n} \sum_{j=1}^{n} (x_j - \overline{x})(y_j - \overline{y}).$$

すると

$$\begin{aligned}
\sum_{j=1}^{n} (y_j - a - bx_j)^2 &= \sum_{j=1}^{n} (y_j - \overline{y} - bx_j + b\overline{x} + \overline{y} - b\overline{x} - a)^2 \\
&= \sum_{j=1}^{n} (y_j - \overline{y} - bx_j + b\overline{x})^2 + 2 \sum_{j=1}^{n} (y_j - \overline{y} - bx_j + b\overline{x})(\overline{y} - b\overline{x} - a) \\
&\quad + n(\overline{y} - b\overline{x} - a)^2 \\
&= \sum_{j=1}^{n} \{ (y_j - \overline{y})^2 - 2b(y_j - \overline{y})(bx_j - b\overline{x}) + (bx_j - b\overline{x})^2 \} + n(\overline{y} - b\overline{x} - a)^2 \\
&= nS_y^2 - 2bnS_{xy} + b^2 nS_x^2 + n(\overline{y} - b\overline{x} - a)^2 \\
&= nS_x^2 \left(b^2 - 2b \frac{S_{xy}}{S_x^2} \right) + nS_y^2 + n(\overline{y} - b\overline{x} - a)^2 \\
&= nS_x^2 \left(b - \frac{S_{xy}}{S_x^2} \right)^2 - n \frac{S_{xy}^2}{S_x^2} + nS_y^2 + n(\overline{y} - b\overline{x} - a)^2.
\end{aligned}$$

これから

$$b = \frac{S_{xy}}{S_x^2}, \quad a = \overline{y} - b\overline{x}$$

のとき最小となる．よって回帰係数 a, b の推定は

$$\hat{b} = \frac{S_{xY}}{S_x^2}, \quad \hat{a} = \overline{Y} - \hat{b}\overline{x} \tag{1.1}$$

で行えばよい．ここで

$$S_{xY} = \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x})(Y_j - \overline{Y}).$$

推定量の分布

次に，区間推定や検定を行うには $\hat{a}, \hat{b}$ の分布を知る必要がある．それに対しては次のことが成り立つ．

定理 1.1. $(\hat{a}, \hat{b})$ は 2 次元正規分布に従い，平均と共分散は次で与えられる：

$$\begin{aligned} E[\hat{a}] &= a, \\ E[\hat{b}] &= b, \\ V(\hat{a}) &= \frac{\sigma^2}{n}\left(1 + \frac{\overline{x}^2}{S_x^2}\right), \\ V(\hat{b}) &= \frac{\sigma^2}{nS_x^2} \\ Cov(\hat{a}, \hat{b}) &= -\frac{\sigma^2\overline{x}}{nS_x^2} \end{aligned}$$

証明　$Y_j = a + bx_j + \varepsilon_j$ であったから，$\overline{Y} = a + b\overline{x} + \overline{\varepsilon}$.

$$\begin{aligned} \hat{b} &= \frac{\sum_j (x_j - \overline{x})(Y_j - \overline{Y})}{nS_x^2} \\ &= \frac{\sum_j (x_j - \overline{x})(a + bx_j + \varepsilon_j - a - b\overline{x} - \overline{\varepsilon})}{nS_x^2} \\ &= \frac{\sum_j b(x_j - \overline{x})(x_j - \overline{x})}{nS_x^2} + \frac{\sum_j (x_j - \overline{x})\varepsilon_j}{nS_x^2} - \frac{\sum_j (x_j - \overline{x})\overline{\varepsilon}}{nS_x^2} \\ &= b + \frac{\sum_j (x_j - \overline{x})\varepsilon_j}{nS_x^2} \\ \hat{a} &= \overline{Y} - \hat{b}\overline{x} \\ &= a + b\overline{x} + \overline{\varepsilon} - \left(b + \frac{\sum_j (x_j - \overline{x})\varepsilon_j}{nS_x^2}\right)\overline{x} \\ &= a + \sum_j \left\{\frac{1}{n} - \frac{\overline{x}(x_j - \overline{x})}{nS_x^2}\right\}\varepsilon_j \\ &= a + \frac{1}{n}\sum_j \left\{1 - \frac{\overline{x}(x_j - \overline{x})}{S_x^2}\right\}\varepsilon_j. \end{aligned}$$

$\hat{a}, \hat{b}$ は独立な正規分布に従う ε_j で表されているから，2 次元正規分布に従う．平均，共分散を計算すると，

$$E[\hat{b}] = b,$$

$$V(\hat{b}) = \frac{\sum (x_j - \overline{x})^2 \sigma^2}{n^2 S_x^4} = \frac{n S_x^2 \sigma^2}{n^2 S_x^4} = \frac{\sigma^2}{n S_x^2}$$

および

$$\begin{aligned}
E[\hat{a}] &= a, \\
V(\hat{a}) &= \frac{1}{n^2} \sum_j \left\{ 1 - \frac{\overline{x}(x_j - \overline{x})}{S_x^2} \right\}^2 \sigma^2 \\
&= \frac{1}{n^2} \sum_j \left\{ 1 - \frac{2\overline{x}(x_j - \overline{x})}{S_x^2} + \frac{\overline{x}^2 (x_j - \overline{x})^2}{S_x^4} \right\} \sigma^2 \\
&= \frac{1}{n^2} \left\{ n + \frac{\overline{x}^2 n S_x^2}{S_x^4} \right\} \sigma^2 \\
&= \frac{\sigma^2}{n} \left(1 + \frac{\overline{x}^2}{S_x^2} \right), \\
Cov(\hat{a}, \hat{b}) &= \frac{1}{n^2 S_x^2} \sum_j (x_j - \overline{x}) \left\{ 1 - \frac{\overline{x}(x_j - \overline{x})}{S_x^2} \right\} \\
&= -\frac{1}{n^2 S_x^2} \sum_j \frac{\overline{x}(x_j - \overline{x})^2}{S_x^2} \\
&= -\frac{1}{n^2 S_x^2} \frac{\overline{x} n S_x^2}{S_x^2} \\
&= -\frac{\sigma^2 \overline{x}}{n S_x^2}
\end{aligned}$$

となる. □

この定理により

$$\frac{\hat{a} - a}{\sqrt{V(\hat{a})}} = \frac{\sqrt{n}(\hat{a} - a)}{\sigma \sqrt{1 + (\overline{x}/S_x)^2}} \sim N(0, 1) \tag{1.2}$$

$$\frac{\hat{b} - b}{\sqrt{V(\hat{b})}} = \frac{\sqrt{n}(\hat{b} - b)}{\sigma / S_x} \sim N(0, 1) \tag{1.3}$$

が分かるから，これを利用して (区間) 推定，検定を行えばよい.

分散の推定

σ^2 が既知の場合はこれでよいが，σ^2 が未知の場合はさらに σ^2 の推定をする必要がある. そのために

$$e_j = Y_j - \hat{b} x_j - \hat{a} = a + b x_j + \varepsilon_j - \hat{b} x_j - \hat{a} = a - \hat{a} + (b - \hat{b}) x_j + \varepsilon_j \tag{1.4}$$

とおく. これは**残差**と呼ばれる. 分散 σ^2 の推定は

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_j e_j^2 \tag{1.5}$$

で行う．この分布が必要であるが，それは次の定理から得られる．

定理 1.2. 残差 $(e_1, e_2, \ldots, e_n)$ は $(\hat{a}, \hat{b})$ とは独立であり，さらに (1.5) で定まる $\hat{\sigma}^2$ も $(\hat{a}, \hat{b})$ と独立で，$\sum_j e_j^2/\sigma^2 = (n-2)\hat{\sigma}^2/\sigma^2$ は自由度 $n-2$ の χ^2 分布に従う．また

$$E[\hat{\sigma}^2] = \sigma^2 \tag{1.6}$$

が成り立ち，$\hat{\sigma}^2$ は σ^2 の不偏推定量である．

証明　すべての確率変数は ε_j で表されているから，正規分布に従うことは明らか．共分散が 0 となることを示せば独立性が示せる．

$$\begin{aligned}
Cov(e_j, \hat{a}) &= Cov(\varepsilon_j - (\hat{a}-a) - (\hat{b}-b)x_j), \hat{a}) \\
&= Cov(\varepsilon_j, \hat{a}) - Cov(\hat{a}-a, \hat{a}) - x_j\, Cov(\hat{b}-b, \hat{a}) \\
&= Cov(\varepsilon_j, \hat{a}) - Cov(\hat{a}, \hat{a}) - x_j\, Cov(\hat{b}, \hat{a}) \\
&= \frac{\sigma^2}{n}\left(1 - \frac{\overline{x}(x_j - \overline{x})}{S_x^2}\right) - \frac{\sigma^2}{n}\left(1 + \frac{\overline{x}^2}{S_x^2}\right) + x_j \frac{\sigma^2 \overline{x}}{nS_x^2} \\
&= \frac{\sigma^2}{n}\left(1 - \frac{\overline{x}(x_j - \overline{x})}{S_x^2} - 1 - \frac{\overline{x}^2}{S_x^2} + \frac{x_j \overline{x}}{S_x^2}\right) \\
&= 0.
\end{aligned}$$

同様に

$$\begin{aligned}
Cov(e_j, \hat{b}) &= Cov(\varepsilon_j - (\hat{a}-a) - (\hat{b}-b)x_j), \hat{b}) \\
&= Cov(\varepsilon_j, \hat{b}) - Cov(\hat{a}-a, \hat{b}) - x_j\, Cov(\hat{b}-b, \hat{b}) \\
&= Cov(\varepsilon_j, \hat{b}) - Cov(\hat{a}, \hat{b}) - x_j\, Cov(\hat{b}, \hat{b}) \\
&= \frac{(x_j - \overline{x})\sigma^2}{nS_x^2} + \frac{\overline{x}\sigma^2}{nS_x^2} - x_j \frac{\sigma^2}{nS_x^2} \\
&= \frac{\sigma^2}{nS_x^2}(x_j - \overline{x} + \overline{x} - x_j) \\
&= 0.
\end{aligned}$$

これで独立性が示せた．

さて，$\hat{a}, \hat{b}$ の定義 (1.1) から

$$\hat{a} + x_j\hat{b} = \overline{Y} - \hat{b}\overline{x} + x_j\hat{b} = \overline{Y} + \frac{S_{xY}}{S_x^2}(x_j - \overline{x}).$$

従って両辺から $a + bx_j$ を引いて

$$\begin{aligned}
\hat{a} - a + (\hat{b}-b)x_j &= \overline{Y} + \frac{S_{xY}}{S_x^2}(x_j - \overline{x}) - a - bx_j \\
&= \overline{Y} - a - b\overline{x} + \frac{S_{xY}}{S_x^2}(x_j - \overline{x}) - b(x_j - \overline{x}) \\
&= \overline{Y} - a - b\overline{x} + \frac{S_{xY} - bS_x^2}{S_x^2}(x_j - \overline{x})
\end{aligned} \tag{1.7}$$

$\overline{Y}$ と S_{xY} について平均，分散を計算する．$Y_j = a + bx_j + \varepsilon_j$ だから Y_1, Y_2,..., Y_n は独立．$\overline{Y} = \frac{1}{n}\sum_j Y_j$ は独立確率変数の和だから

$$E[\overline{Y}] = \frac{1}{n}\sum_j E[Y_j] = \frac{1}{n}\sum_j (a + bx_j) = a + b\overline{x},$$
$$V(\overline{Y}) = \sum_j V(\frac{Y_j}{n}) = \frac{1}{n^2}\sum_j V(Y_j) = \frac{\sigma^2}{n}.$$

S_{xY} については

$$S_{xY} = \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x})(Y_j - \overline{Y}) = \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x})Y_j$$

とやはり独立確率変数の和だから

$$\begin{aligned}
E[S_{xY}] &= \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x})E[Y_j] = \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x})(a + bx_j) \\
&= \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x})bx_j = \frac{b}{n}\sum_{j=1}^{n}(x_j - \overline{x})(x_j - \overline{x}) \\
&= bS_x^2, \\
V(S_{xY}) &= \frac{1}{n^2}\sum_j V((x_j - \overline{x})Y_j) = \frac{1}{n^2}\sum_j (x_j - \overline{x})^2 V(Y_j) = \frac{1}{n^2}\sum_j (x_j - \overline{x})^2\sigma^2 \\
&= \frac{\sigma^2 S_x^2}{n}, \\
Cov(\overline{Y}, S_{xY}) &= \frac{1}{n^2}\sum_j (x_j - \overline{x})V(Y_j) = \frac{1}{n^2}\sum_j (x_j - \overline{x})\sigma^2 \\
&= 0.
\end{aligned}$$

共分散が 0 だから独立であることが分かる．さらに

$$\frac{\sqrt{n}(\overline{Y} - a - b\overline{x})}{\sigma} \sim N(0,1), \quad \frac{\sqrt{n}(S_{xY} - bS_x^2)}{\sigma S_x} \sim N(0,1)$$

であるから 2 乗して加えれば,

$$\frac{n(\overline{Y} - a - b\overline{x})^2}{\sigma^2} + \frac{n(S_{xY} - bS_x^2)^2}{\sigma^2 S_x^2} \sim \chi^2(2) \tag{1.8}$$

であることが分かる．

これは予測誤差の 2 乗和（を σ^2 で割ったもの）といわれるものである．そのことを説明しておこう．$\varepsilon_j = Y_j - (a + bx_j)$ は観測誤差と呼ばれるが，(1.4) の残差の関係式 $e_j = a - \hat{a} + (b - \hat{b})x_j + \varepsilon_j$ を用いて次の様に分解される

$$\varepsilon_j = \underbrace{Y_j - (a + bx_j)}_{\text{観測誤差}} = \underbrace{\hat{a} - a + (\hat{b} - b)x_j}_{\text{予測誤差}} + \underbrace{e_j}_{\text{残差}}$$

ここで (1.7) を用いて予測誤差は

$$\hat{a} - a + (\hat{b} - b)x_j = \overline{Y} - a - b\overline{x} + \frac{S_{xY} - bS_x^2}{S_x^2}(x_j - \overline{x})$$

であるからその 2 乗和は

$$\begin{aligned}
&\sum_j (\hat{a} - a + (\hat{b} - b)x_j)^2 \\
&\quad = \sum_j \left\{ \overline{Y} - a - b\overline{x} + \frac{S_{xY} - bS_x^2}{S_x^2}(x_j - \overline{x}) \right\}^2 \\
&\quad = \sum_j \left\{ (\overline{Y} - a - b\overline{x})^2 + 2(\overline{Y} - a - b\overline{x})\frac{S_{xY} - bS_x^2}{S_x^2}(x_j - \overline{x}) + \frac{(S_{xY} - bS_x^2)^2}{S_x^4}(x_j - \overline{x})^2 \right\} \\
&\quad = n(\overline{Y} - a - b\overline{x})^2 + \frac{(S_{xY} - bS_x^2)^2}{S_x^4} n S_x^2 \\
&\quad = n(\overline{Y} - a - b\overline{x})^2 + \frac{n(S_{xY} - bS_x^2)^2}{S_x 2}.
\end{aligned}$$

これを σ^2 で割ったものが，(1.8) の左辺であった．これから次のことも分かる．

$$E[\sum_j (\hat{a} - a + (\hat{b} - b)x_j)^2] = 2\sigma^2. \tag{1.9}$$

さらに，予測誤差と残差の関係をもう少し見ておこう．まず，

$$\overline{e} = \frac{1}{n}\sum_{j=1}^{n} e_j = 0 \tag{1.10}$$

に注意しよう．これは次のようにしてわかる．

$$e_j = Y_j - \hat{a} - \hat{b}x_j$$

より

$$\overline{e} = \overline{Y} - \hat{a} - \hat{b}\overline{x} = 0. \quad (\because (1.1) \text{ より } \hat{a} = \overline{Y} - \hat{b}\overline{x})$$

これを使って

$$\sum_{j=1}^{n} e_j(\hat{a} - a + (\hat{b} - b)x_j) = n(\hat{a} - a)\overline{e} + (\hat{b} - b)\sum_{j=1} e_j x_j = (\hat{b} - b)\sum_{j=1} e_j x_j.$$

さらに

$$\sum_{j=1}^{n} e_j x_j = \sum_{j=1}^{n} (Y_j - \hat{b}x_j - \hat{a})x_j = \sum_{j=1}^{n} (Y_j - \overline{Y} - \hat{b}x_j + \hat{b}\overline{x} + \overline{Y} - \hat{b}\overline{x} - \hat{a})x_j$$

$$
\begin{aligned}
&= \sum_{j=1}^{n}(Y_j - \overline{Y})x_j - \hat{b}\sum_{j=1}^{n}(x_j - \overline{x})x_j \\
&= \sum_{j=1}^{n}(Y_j - \overline{Y})(x_j - \overline{x}) - \hat{b}\sum_{j=1}^{n}(x_j - \overline{x})(x_j - \overline{x}) \\
&= nS_{xY} - \frac{S_{xY}}{S_x^2}nS_x^2 = 0.
\end{aligned}
$$

結局

$$\sum_{j=1}^{n} e_j(\hat{a} - a + (\hat{b} - b)x_j) = 0.$$

一方

$$\varepsilon_j = e_j + (\hat{a} - a + (\hat{b} - b)x_j)$$

だから

$$\sum_j \varepsilon_j^2 = \sum_j e_j^2 + \sum_j (\hat{a} - a + (\hat{b} - b)x_j)^2$$

これで

$$\chi^2(n) \qquad\qquad \chi^2(2)$$

$$\text{観測誤差の 2 乗和} \quad = \quad \text{残差の 2 乗和} \quad + \quad \text{予測誤差の 2 乗和}$$

であることが分かった．(観測誤差の 2 乗和)$/\sigma^2 \sim \chi^2(n)$ であり (予測誤差の 2 乗和)$/\sigma^2 \sim \chi^2(2)$ である．さらに予測誤差の 2 乗和と残差の 2 乗和は独立である．このことは $(\hat{a}, \hat{b})$ と e_j が独立であることから分かる．従って (残差の 2 乗和)$/\sigma^2 \sim \chi^2(n-2)$ であることが分かる．

また

$$E[\sum_j \varepsilon_j^2] = n\sigma^2$$

と (1.9) から

$$E[\sum_j e_j^2] = (n-2)\sigma^2$$

が分かる．以上で主張が従う． □

上の定理から残差 2 乗和を $n-2$ で割った $\hat{\sigma}^2$ が σ^2 の不偏推定量であることが分かり，(1.2), (1.3) 式の σ^2 を $\hat{\sigma}^2$ で置き換えた

$$T_a = \frac{\sqrt{n}(\hat{a} - a)}{\hat{\sigma}\sqrt{1 + (\overline{x}/S_x)^2}} \sim t(n-2), \tag{1.11}$$

$$T_b = \frac{\sqrt{n}(\hat{b} - b)}{\hat{\sigma}/S_x} \sim t(n-2) \tag{1.12}$$

で a, b の推定，検定を行う．$\hat{\sigma}^2/\sigma^2 \sim \chi^2(n-2)/(n-2)$ であったから，T_a, T_b が 自由度 $n-2$ の t 分布に従うことも分かったことになる．

2. 分散分析

等平均の検定を 第 4 章 第 2 節 で行ったが，これは 2 標本問題であった．3 つ以上の標本を扱う場合は多標本問題となり，さらに違った方法が必要となる．この問題は，都市と所得との関係を考える場合などに現れてくる．都市部の水準として，人口により 10 万，50 万，100 万，500 万などと分けることにより，多標本問題となるわけである．

1 元間配置

ここではもっとも単純な一元配置の問題のみを扱う．要因 A を r 水準 (level) $A_1, A_2, \dots A_r$ に分類し，各水準における標本を

$$\begin{aligned} A_1 \text{ 標本} \quad & Y_{11}, \dots, Y_{1n_1} \sim N(\mu_1, \sigma^2) \\ & \vdots \\ A_r \text{ 標本} \quad & Y_{r1}, \dots, Y_{rn_r} \sim N(\mu_r, \sigma^2) \end{aligned}$$

とし，各水準は独立であるとする．このとき，母平均 $\mu_1, \dots, \mu_r$ の間に差があるかどうか検定しよう．次のように仮説を立てる．

$$\begin{cases} H_0 : \ \mu_1 = \cdots = \mu_r \\ H_1 : \ \mu_1, \dots, \mu_r \text{ のどれかが異なる} \end{cases}$$

総標本数を $n = n_1 + \cdots + n_r$ とすると，総平均と各平均偏差は

$$\mu = \frac{1}{n} \sum_{i=1}^{r} n_i \mu_i, \quad \alpha_i = \mu_i - \mu$$

となる．ここで $\sum_{i=1}^{r} n_i \alpha_i = 0$ に注意しておこう．α_i は水準 A_i の**主効果** (main effect) と呼ばれる．これを用いれば仮説は次のようにも表現できる．

$$\begin{cases} H_0 : \ \alpha_1 = \cdots = \alpha_r = 0 \\ H_1 : \ \alpha_1, \dots, \alpha_r \text{ のどれかが } 0 \text{ と異なる} \end{cases}$$

以下しばらく，帰無仮説を仮定しないで計算を進める．観測は

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1, \dots, r \,;\ j = 1, \dots, n_i$$

と表現される．ここで ε_{ij} は誤差を表し，$N(0, \sigma^2)$ に従う独立確率変数である．

推定量

次のような統計量を考える.

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} Y_{ij} \sim N(\mu, \frac{1}{n}\sigma^2)$$
$$\overline{Y}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} Y_{1j} \sim N(\mu_1, \frac{1}{n_1}\sigma^2)$$
$$\vdots$$
$$\overline{Y}_r = \frac{1}{n_r}\sum_{j=1}^{n_r} Y_{rj} \sim N(\mu_r, \frac{1}{n_r}\sigma^2).$$

次が成立していることに注意しよう.

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{r} n_i \overline{Y}_i. \tag{2.1}$$

ここで

$$\hat{\alpha}_i = \overline{Y}_i - \overline{Y} \sim N(\alpha_i, (\frac{1}{n_i} - \frac{1}{n})\sigma^2)$$

とおく. $\sum_{i=1}^{r} n_i\hat{\alpha}_i = 0$ が成り立っている. 誤差の 2 乗和を計算すると

$$\begin{aligned}
\Delta^2 &= \sum_{i=1}^{r}\sum_{j=1}^{n_i} \varepsilon_{ij}^2 \\
&= \sum_{i=1}^{r}\sum_{j=1}^{n_i} (Y_{ij} - \mu - \alpha_i)^2 \\
&= \sum_{i=1}^{r}\sum_{j=1}^{n_i} \{(Y_{ij} - \overline{Y} - \hat{\alpha}_i) + (\overline{Y} - \mu) + (\hat{\alpha}_i - \alpha_i)\}^2 \\
&= \underbrace{\sum_{i=1}^{r}\sum_{j=1}^{n_i} (Y_{ij} - \overline{Y} - \hat{\alpha}_i)^2}_{} + \underbrace{n(\overline{Y} - \mu)^2 + \sum_{i=1}^{r} n_i(\hat{\alpha}_i - \alpha_i)^2}_{\text{最尤法でここを } 0 \text{ にするのが最良}}
\end{aligned}$$

となる. ここで交差項が消えるのは

$$\sum_{j=1}^{n_i} (Y_{ij} - \overline{Y} - \hat{\alpha}_i) = n_i(\overline{Y}_i - \overline{Y} - \hat{\alpha}_i) = 0$$

および

$$\sum_{i=1}^{r} n_i(\hat{\alpha}_i - \alpha_i) = 0$$

による．上の計算から Δ^2 を最小にするには

$$\mu = \overline{Y}, \quad \alpha_i = \hat{\alpha}_i$$

とすればよいことが分かる．すなわち μ と α_i の推定量が $\overline{Y}, \hat{\alpha}_i$ であることが，最小二乗法によって確かめられたことになる．正規分布であるから，最小二乗法による方法は最尤法とも一致する．

残差平方和は

$$\Delta_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y} - \hat{\alpha}_i)^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2$$

である．これから次のような変動量を定義する:

級内変動

$$\Delta_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2.$$

総変動

$$\Delta_1^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y})^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y} - \hat{\alpha}_i)^2 + \sum_{i=1}^{r} n_i\hat{\alpha}_i^2.$$

級間変動

$$\Delta_1^2 - \Delta_0^2 = \sum_{i=1}^{r} n_i\hat{\alpha}_i^2 = \sum_{i=1}^{r} n_i(\overline{Y}_i - \overline{Y})^2.$$

級内変動 Δ_0^2 の自由度は観測した個数 n と，推定した個数 r (μ_i を $\overline{Y}_i$ で推定している) の差 $n - r$ である．これはまた Δ_0^2 の平均をとって

$$\begin{aligned}
E[\Delta_0^2] &= \sum_{i=1}^{r}\sum_{j=1}^{n_i} E[(Y_{ij} - \overline{Y}_i)^2] \\
&= \sum_{i=1}^{r}\sum_{j=1}^{n_i} E[\{Y_{ij} - \mu_i - (\overline{Y}_i - \mu_i)\}^2] \\
&= \sum_{i=1}^{r}\sum_{j=1}^{n_i} E[(Y_{ij} - \mu_i)^2] - 2\sum_{i=1}^{r}\sum_{j=1}^{n_i} E[(Y_{ij} - \mu_i)(\overline{Y}_i - \mu_i)] + \sum_{i=1}^{r}\sum_{j=1}^{n_i} E[(\overline{Y}_i - \mu_i)^2] \\
&= \sum_{i=1}^{r}\sum_{j=1}^{n_i} E[(Y_{ij} - \mu_i)^2] - 2\sum_{i=1}^{r} n_i E[(\overline{Y}_i - \mu_i)(\overline{Y}_i - \mu_i)] + \sum_{i=1}^{r} n_i E[(\overline{Y}_i - \mu_i)^2] \\
&= \sum_{i=1}^{r}\sum_{j=1}^{n_i} E[(Y_{ij} - \mu_i)^2] - \sum_{i=1}^{r} n_i E[(\overline{Y}_i - \mu_i)^2]
\end{aligned}$$

$$= n\sigma^2 - \sum_{i=1}^{r} n_i \frac{\sigma^2}{n_i}$$
$$= n\sigma^2 - r\sigma^2 = (n-r)\sigma^2$$

であることからも確かめられる．$Y_{ij} - \overline{Y}_i$ の平均は 0 であるから，Δ_0^2/σ^2 の分布は $\chi^2(n-r)$ であることが分かる．この場合は帰無仮説を仮定しなくても成り立っている．

総変動は Δ_1^2 の自由度は観測した個数 n と，推定した個数 1 (μ を $\overline{Y}$ で推定している) の差 $n-1$ である．

級間変動に関しては，自由度は $(n-1)-(n-r) = r-1$ である．これはまた $\Delta_1^2 - \Delta_0^2$ の平均をとって

$$\begin{aligned}
E[\Delta_1^2 - \Delta_0^2] &= \sum_{i=1}^{r} n_i E[(\overline{Y}_i - \overline{Y})^2] \\
&= \sum_{i=1}^{r} n_i E[\overline{Y}_i^2] - 2\sum_{i=1}^{r} n_i E[\overline{Y}_i \overline{Y}] + \sum_{i=1}^{r} n_i E[\overline{Y}^2] \\
&= \sum_{i=1}^{r} n_i E[\overline{Y}_i^2] - 2nE[\overline{Y}\overline{Y}] + nE[\overline{Y}^2] \\
&= \sum_{i=1}^{r} n_i E[\overline{Y}_i^2] - nE[\overline{Y}^2] \\
&= \sum_{i=1}^{r} n_i \frac{\sigma^2}{n_i} - n\frac{\sigma^2}{n} \\
&= (r-1)\sigma^2
\end{aligned}$$

となることからも確かめられる．級間変動に関しては帰無仮説を仮定すると $\overline{Y}_i - \overline{Y}$ の平均が 0 であるから $(\Delta_1^2 - \Delta_0^2)/\sigma^2$ の分布は $\chi^2(r-1)$ となる．

級内変動と級間変動は独立になっている．これは共分散を計算して

$$\begin{aligned}
E[(Y_{ij} - \overline{Y}_i)(\overline{Y}_k - \mu_k - \overline{Y} + \mu)] &= E[(Y_{ij} - \mu_i - \overline{Y}_i + \mu_i)(\overline{Y}_k \mu_k - \overline{Y} + \mu)] \\
&= E[(Y_{ij} - \mu_i)(\overline{Y}_k \mu_k)] - E[(Y_{ij} - \mu_i)(\overline{Y} - \mu)] \\
&\quad - E[(\overline{Y}_i - \mu_i)(\overline{Y}_k \mu_k)] + E[(\overline{Y}_i - \mu_i)(\overline{Y} - \mu)] \\
&= \delta_{ik}\frac{\sigma^2}{n_i} - \frac{\sigma^2}{n} - \delta_{ik}\frac{\sigma^2}{n_i} + \frac{n_i}{n}\frac{\sigma^2}{n_i} = 0
\end{aligned}$$

となることから分かる．

以上自由度について纏めると

$$\begin{aligned}
&\text{変動：} \quad \text{総変動} = \text{級間変動} + \text{級内変動} \\
&\text{自由度：} \quad n-1 \;=\; r-1 \;+\; n-r
\end{aligned}$$

ここで次の統計量を定義する

$$F = \frac{\dfrac{\Delta_1^2 - \Delta_0^2}{r-1}}{\dfrac{\Delta_0^2}{n-r}}.$$

すると帰無仮説の下では $(\Delta_1^2 - \Delta_0^2)/\sigma^2$ は自由度 $r-1$ の χ^2 分布に従い，Δ_0^2/σ^2 は 自由度 $n-r$ の χ^2 分布に従う．また独立性も成立しているので上の F は自由度 $(r-1, n-r)$ の F-分布に従う．このことを利用して，有意水準 α を与えて棄却域を

$$W = \{f : f > F_\alpha(n-r, r-1)\}$$

と定めて，検定を行うことになる．このような分析方法を**分散分析** (analysis of variance = ANOVA) という．

これを整理して，次の表が得られる．この表を分散分析表という

分散分析表 (SS ＝平方和，DF ＝自由度，MS ＝平均平方和)

	SS	DF	MS	F
級間	$\sum n_i(\overline{Y}_i - \overline{Y})^2$	$r-1$	$\dfrac{\sum n_i(\overline{Y}_i - \overline{Y})^2}{r-1}$	$F = \dfrac{\dfrac{\sum n_i(Y_{ij} - \overline{Y}_i)^2}{r-1}}{\dfrac{\sum\sum(Y_{ij} - \overline{Y}_i)^2}{n-r}}$
級内	$\sum\sum(Y_{ij} - \overline{Y}_i)^2$	$n-r$	$\dfrac{\sum\sum(Y_{ij} - \overline{Y}_i)^2}{n-r}$	
総	$\sum\sum(Y_{ij} - \overline{Y})^2$	$n-1$		

関連図書

[1] 松原望, 縄田和満, 中井検裕, 統計学入門, 東京大学出版会, 東京, 1991.

[2] R. V. Hogg and A. T. Craig,“*Introduction to Mathematical Statistics*,” Macmillan Company, London, 1970.

[3] 稲垣宣生, 数理統計学, 改訂版, 数学シリーズ, 裳華房, 東京, 2003.

[4] 河田 敬義, 丸山 文行, 鍋谷 清治, 大学演習 数理統計, 裳華房, 東京, 1962.

[5] 国沢 清典 編, 確率統計演習 2 統計, 培風館, 東京, 1966.

[6] 吉田 伸生, 確率の基礎から統計へ, 遊星社, 東京, 2012.