標本分布

数学と医療・第 7 回

数学・橋本貴宏

2024 年 10 月 16 日 1 限

本日の授業では

- 前回のおさらい
  - (1) 母集団と標本
  - (2) 標本平均の分布
- 標本分布
  - (1) χ^2 分布
  - (2) t 分布
  - (3) F 分布

前回のおさらい (その1)

母集団: 調査, 観察の対象となる集団

標本: 母集団の調査のために取り出したデータ.

標本抽出は**無作為抽出** (ランダムに選びだすということ) によってなされる.

標本平均: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ $\cdots$ 標本を取り出すごとに変わる確率変数である.

標本分散: $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$, **不偏分散**: $U^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

$\overline{X}$ の分布: 母平均 μ, 母分散 σ^2 の (無限) 母集団から抽出された大きさ n の標本 X_i ($i = 1, \cdots, n$) の標本平均 $\overline{X}$ について, $E(\overline{X}) = \mu$, $V(\overline{X}) = \frac{\sigma^2}{n}$ が成り立つ.

前回のおさらい (その 2)

正規母集団からの標本平均の分布

正規母集団 $\boldsymbol{N(\mu, \sigma^2)}$ から抽出された大きさ $\boldsymbol{n}$ の標本の標本平均 $\boldsymbol{\overline{X}}$ は,
$\boldsymbol{N(\mu, \frac{\sigma^2}{n})}$ に従う.

中心極限定理:　母平均 $\boldsymbol{\mu}$, 母分散 $\boldsymbol{\sigma^2}$ の無限母集団から大きさ $\boldsymbol{n}$ の標本を独立に無作為抽出する. $\boldsymbol{\overline{X}}$ は $\boldsymbol{n}$ が十分大きいとき $\boldsymbol{N(\mu, \frac{\sigma^2}{n})}$ で近似できる.

(これの二項分布版は「ド・モアブル-ラプラスの定理」と呼ばれる: $\boldsymbol{X}$ が $\boldsymbol{B(n, p)}$ に従うとき, $\boldsymbol{Z = \frac{X - np}{\sqrt{np(1-p)}}}$ と標準化すると, $\boldsymbol{n \to \infty}$ のとき $\boldsymbol{Z}$ は $\boldsymbol{N(0, 1)}$ に従う.)

標本比率の分布:　母比率 $\boldsymbol{p}$ の二項母集団から, 大きさ $\boldsymbol{n}$ の標本を抽出するとき, 標本比率 $\boldsymbol{\widehat{P}}$ は $\boldsymbol{n}$ が十分大きいならば, 近似的に $\boldsymbol{N(p, \frac{p(1-p)}{n})}$ に従う.

前回の演習問題

問題

令和 4 年度学校保健統計調査 (文部科学省) によると, ある県の 17 歳女子の身長は, 平均 $\mathbf{158.0}$ (cm), 標準偏差 $\mathbf{5.70}$ (cm) の正規分布に従うという。次の問いに答えよ。

(1) この中から無作為抽出で選んだ $\mathbf{15}$ 人の身長の平均を $\overline{\boldsymbol{X}}$ (cm) とするとき, $\overline{\boldsymbol{X}}$ はどのような確率分布に従うか。

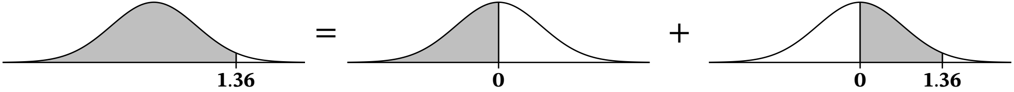
(2) $\overline{\boldsymbol{X}}$ が $\mathbf{160.0}$ 以下である確率を求めよ。

演習問題の解説

(1) 母集団は $N(\mu, \sigma^2) = N(158.0, 5.7^2)$ に従っている．$n = 15$ の標本の標本平均 $\overline{X}$ は $N(158, \frac{5.7^2}{15}) = N(158, 2.166)$ に従う．($\frac{5.7}{\sqrt{15}} \fallingdotseq 1.4717\cdots$ であるから，平均 158.0, 標準偏差 1.47 の正規分布に従う)

(2) $\overline{X}$ の標準化を Z とすると，$Z = \dfrac{\overline{X} - 158}{\sqrt{2.166}}$．よって求める確率は

$$P(\overline{X} \leqq 160) = P\left(Z \leqq \frac{160 - 158}{\sqrt{2.166}}\right) \fallingdotseq P(Z \leqq 1.36) = 0.5 + 0.4131 = 0.9131.$$



みなさんからのコメント

- 演習問題を解く紙があってありがたいです
  数式を手書きで書くためには, 紙の方が良いのでしょう
- 今回の内容は前回の授業内容に比べて、難易度が上がったように感じました。数式などを用いて説明しているところは難しかったが, 標本サイズを増やすことで、誤差が減るということは理解できました
  この点は大事なところです．分散が $\frac{\sigma^2}{n}$ なので, n が大きいほどバラツキが小さくなります
- 買った関数電卓が思ったより様々なことに使えるとわかった
  標準偏差とか, 相関係数とかが瞬時に求まるのです
- 正規分布の再生性は要するに、分母が大きくなると誤差が小さくなる。ばらけていたデータがグラフの形になっていくということなのかなと思った
  再生性は, 和をとっても同じ分布になるということです

3 つの標本分布

母集団がどのような確率分布であるか調べるために標本を抽出する．標本から求める統計量を標本統計量という．**標本分布**は推定・検定などの推測統計で用いられる確率分布であり，次の 3 つが代表的である．

(1) χ^2 **分布**

母分散の推定および適合度・独立性の検定に使われる．

(2) t **分布**

正規母集団の母分散が未知のとき，**母平均**の区間推定や検定，2 つの正規母集団の母分散が等しいときの母平均の差の検定に使われる．

(3) F **分布**

2 つの正規母集団における**母分散の**間に**差**があるかどうかを検定するときに使われる．

χ^2 分布 (p.88)

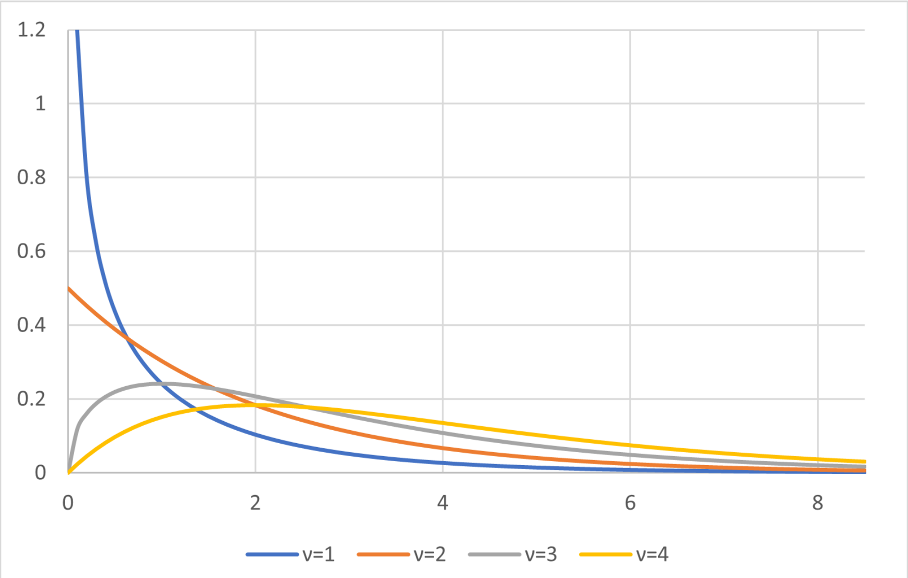
$\boldsymbol{n}$ を自然数とする. 連続型確率変数 $\boldsymbol{X}$ の確率密度関数が

$$\boldsymbol{f_n(x)} = \begin{cases} \dfrac{\mathbf{1}}{\mathbf{2}^{\frac{n}{2}}\Gamma(\frac{n}{2})}\boldsymbol{x}^{\frac{n}{2}-1}\boldsymbol{e}^{-\frac{x}{2}} & (\boldsymbol{x} > \mathbf{0}) \\ \mathbf{0} & (\boldsymbol{x} \leqq \mathbf{0}) \end{cases}$$

で与えられるとき, $\boldsymbol{X}$ の確率分布を自由度 $\boldsymbol{n}$ の**カイ２乗分布** (χ^2 分布) といい, $\chi^2(\boldsymbol{n})$ で表す. ここで, $\boldsymbol{\Gamma(\alpha)}$ は正の数 $\boldsymbol{\alpha}$ に対して $\boldsymbol{\Gamma(\alpha)} = \displaystyle\int_0^\infty \boldsymbol{x^{\alpha-1}e^{-x}dx}$ で定義される関数で, **ガンマ関数**と呼ばれる. (ちなみに $\boldsymbol{\Gamma(k+1) = k!}$ である)

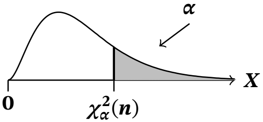
χ^2 分布の確率密度関数のグラフは, 自由度によって大きく形を変える.

χ^2 分布の確率密度関数のグラフ



χ^2 分布表 (p.88)

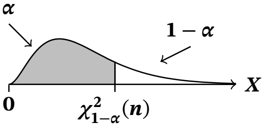
$X \sim \chi^2(n)$ であるとき, $0 \leqq \alpha \leqq 1$ を満たす α に対して $P(X \geqq k) = \alpha$ を満たす点を $\chi^2(n)$ の上側 $100\alpha\%$ 点といい, $\chi^2_\alpha(n)$ で表す. $\chi^2(n)$ の値は p.203 の χ^2 分布表より求める.



例 4.4.1 (1) $\chi^2(12)$ の上側 5% 点は, $\chi^2_{0.05}(12) = 21.026$.
(2) $\chi^2(16)$ の上側 95% 点は, $t_{0.95}(16) = 7.962$.

上側 $100\alpha\%$ 点は $\chi^2_\alpha(n)$ であり,

下側 $100\alpha\%$ の点は $\chi^2_{1-\alpha}(n)$ となる.



χ^2 分布の性質

(1) $X \sim \chi^2(n)$ のとき, $E(X) = n$, $V(X) = 2n$.

(2) χ^2 **分布の再生性**: 確率変数 X と Y が互いに独立で, $X \sim \chi^2(n_1)$, $Y \sim \chi^2(n_2)$ ならば $X + Y \sim \chi^2(n_1 + n_2)$.

(3) $N(0, 1)$ に従う確率変数 Z_1, Z_2, $\cdots$, Z_n が互いに独立ならば
$\displaystyle\sum_{i=1}^{n} Z_i^2 \sim \chi^2(n)$ が成り立つ.

★ これより, 確率変数 X_1, X_2, $\cdots$, X_n が互いに独立で, すべて $N(\mu, \sigma^2)$ に従うとき,
$\displaystyle\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 \sim \chi^2(n)$ であることがわかるので, χ^2 分布はデータ (X_i) と平均 (期待値) とのズレ具合を表している分布であると考えられる.

標本分散, 不偏分散に関する分布 (p.90)

正規母集団からの標本の不偏分散 U^2 (あるいは S^2) は χ^2 分布に従う.

★ このことは, 今後学ぶ区間推定, 仮説検定において重要な役割を果たす.

定理 4.4.2 $X_1, X_2, \cdots, X_n$ を $N(\mu, \sigma^2)$ からの標本とするとき,

(1) 標本平均 $\overline{X}$ と $SS = \sum_{i=1}^{n}(X_i - \overline{X})^2$ (偏差平方和) は独立

(2) $\dfrac{nS^2}{\sigma^2} = \dfrac{(n-1)U^2}{\sigma^2} = \dfrac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \sim \chi^2(n-1)$

- $\dfrac{x_1 + \cdots + x_n}{n} = \bar{X}$ という関係式があるので, 自由度が 1 減っている.

t 分布 (その1) (p.91)

n を自然数とする. 連続型確率変数 X の確率密度関数が

$$f_n(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\,\Gamma(\frac{n}{2})}\left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

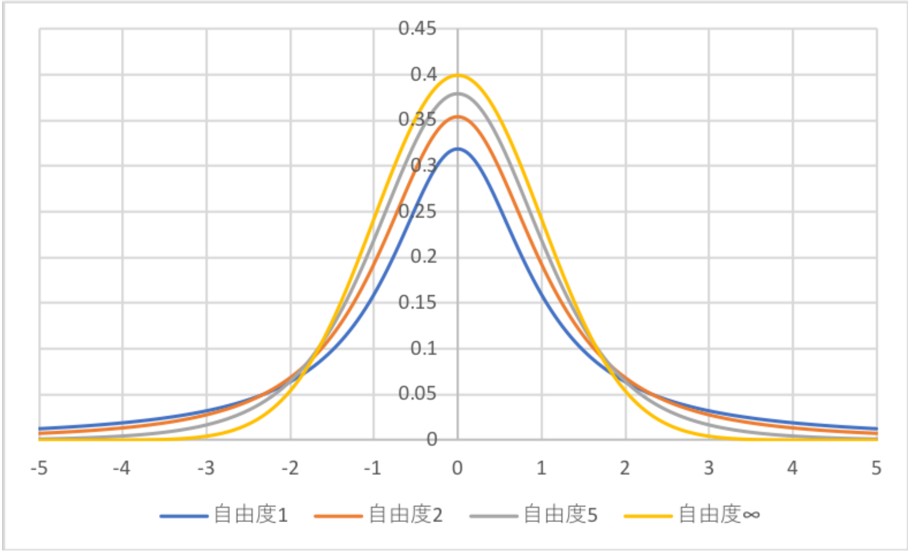
で与えられるとき, X の確率分布を自由度 n の T **分布** といい, $t(n)$ で表す.

に従う確率分布を, **自由度 n の t 分布**という.

t 分布の確率密度関数のグラフ

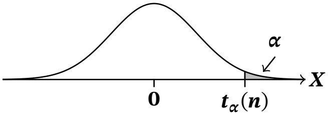
(1) $x = 0$ に関して対称
(2) t 分布は自由度 $n \to \infty$ のとき $N(0, 1)$ に近づく

t 分布の確率密度関数のグラフ (p.81)



t 分布表 (p.92)

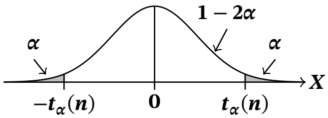
$X \sim t(n)$ であるとき, $0 \leqq \alpha \leqq 1$ を満たす α に対して $P(X \geqq k) = \alpha$ を満たす点を $t(n)$ の上側 $100\alpha\%$ 点といい, $t_\alpha(n)$ で表す. $t(n)$ の値は p.204 の t 分布表より求める.



例 4.5.1　(1) $t(15)$ の上側 5% 点は, $t_{0.05}(15) = 1.753$.

(2) $t(20)$ の上側 2.5% 点は, $t_{0.025}(20) = 2.086$.

t 分布は左右対称であるから, $t(n)$ の上側 $100\alpha\%$ 点 $t_\alpha(n)$ に対し, $-t_\alpha(n)$ は下側 $100\alpha\%$ 点になる.



(3) $t(10)$ の下側 2.5% 点は $-t_{0.025}(10) = -2.228$

定理 4.5.1 (1) X が $t(n)$ に従うとき, $n \geqq 2$ のとき $E(X) = 0$ ($n = 1$ のときは期待値は存在しない), $n \geqq 3$ のとき $V(X) = \dfrac{n}{n-2}$.

(2) 確率変数 X, Y が独立で, $X \sim N(0,1)$, $Y \sim \chi^2(n)$ であるとき,
$T = \dfrac{X}{\sqrt{\frac{Y}{n}}} \sim t(n)$.

正規母集団 $N(\mu, \sigma^2)$ から抽出されら大きさ n の標本の標本平均 $\overline{X}$, 不偏分散 U^2 に対して

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

t 分布の性質 (その 2)

が成り立つ. Z と Y は独立であるから,

$$T = \frac{Z}{\sqrt{Y/(n-1)}} = \frac{\dfrac{\sqrt{n}(\overline{X}-\mu)}{\sigma}}{\sqrt{\dfrac{(n-1)U^2}{\sigma^2(n-1)}}} = \frac{\sqrt{n}(\overline{X}-\mu)}{U}$$

は自由度 $n-1$ の t 分布に従う. (定理 4.5.2)

★ この事実より, t 分布は, (正規母集団の母分散が未知のとき) 母**平均**の区間推定や母平均の検定に利用される.

F 分布 (p.94)

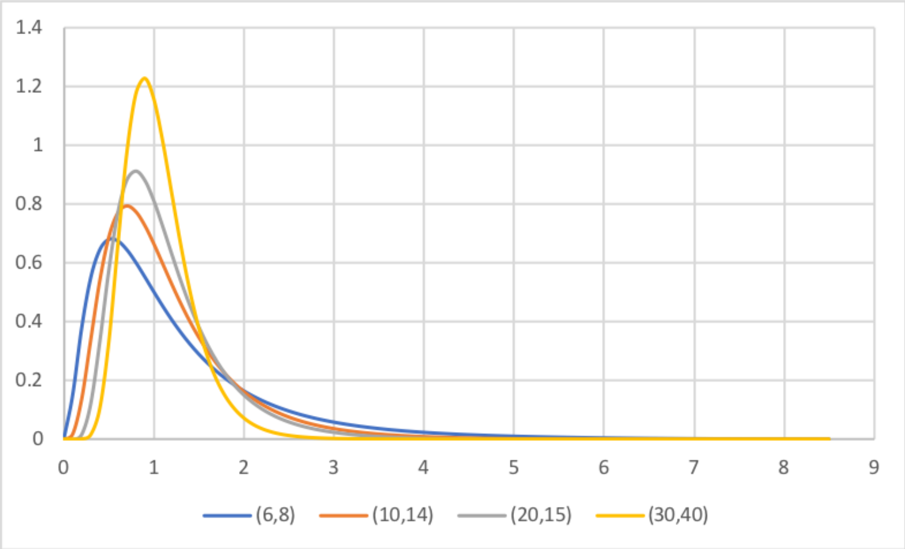
m, n を自然数とする. 連続型確率変数 X の確率密度関数が

$$f_n(x) = \begin{cases} \dfrac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \left(\dfrac{m}{n}\right)^{\frac{m}{2}} x^{\frac{m-2}{2}} \left(1 + \dfrac{m}{n}x\right)^{-\frac{m+n}{2}} & (x > 0), \\ 0 & (x \leqq 0) \end{cases}$$

で与えられるとき, X の確率分布を自由度 (m, n) の **F 分布** といい, $F(m, n)$ で表す.

※ m を第 1 自由度, n を第 2 自由度ということもある. 後述の 定理 4.6.1 (2) において, 分子にあたるのが m, 分母にあたるのが n である.

自由度 (m, n) の F 分布の確率密度関数のグラフ



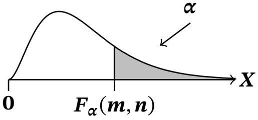
定理 4.6.1 (1) $X \sim F(m, n)$ のとき, $n \geqq 3$ のとき $E(X) = \dfrac{n}{n-2}$, $n \geqq 5$ のとき $V(X) = \dfrac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$ である.

(2) 確率変数 X, Y が互いに独立で, $X \sim \chi^2(m)$, $Y \sim \chi^2(n)$ とするとき,
$F = \dfrac{X/m}{Y/n} \sim F(m, n)$ が成り立つ.

(3) $X \sim F(m, n)$ ならば $\dfrac{1}{X} \sim F(n, m)$ が成り立つ.

F 分布表と読み方 (p.95)

$X \sim F(m,n)$ であるとき, $0 \leqq \alpha \leqq 1$ を満たす α に対して $P(X \geqq k) = \alpha$ を満たす点を $F(m,n)$ の上側 $100\alpha\%$ 点といい, $F_\alpha(m,n)$ で表す



p.205–210 の F 分布表に $\alpha = 0.05,\ 0.025$ と 0.01 の場合の $F_\alpha(m,n)$ の値がある.

例 4.6.1 (1) $F(4,6)$ の上側 5% 点は $F_{0.05}(4,6) = 4.53$

(2) $F(10,10)$ の上側 2.5% 点は $F_{0.025}(10,10) = 3.72$

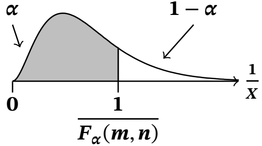
F 分布の上側 $100\alpha\%$ 点について, 定理 4.6.1 (3) より,

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

下側 $100\alpha\%$ 点 $=$ 上側 $100(1-\alpha)\%$ 点の求め方

が成立する.

(証)　$X \sim F(m,n)$ なら,



$$\alpha = P(X \geqq F_\alpha(m,n)) = P\left(\frac{1}{X} \leqq \frac{1}{F_\alpha(m,n)}\right)$$

と書ける. 一方, $\frac{1}{X} \sim F(n,m)$ なので, $F_{1-\alpha}(n,m) = \frac{1}{F_\alpha(m,n)}$ である.

例 4.6.1　(3) $F(3,7)$ の下側 1% 点 $= F(3,7)$ の上側 99% 点は,

$$F_{0.99}(3,7) = \frac{1}{F_{0.01}(7,3)} = \frac{1}{27.7} = 0.036$$

※ Excel では, 上側 α 点 は F.INV.RT(α, m, n), 下側 α 点は F.INV(α, m, n) とすれば求められる.

F 分布と分散比 (p.97)

定理 4.6.2 2 つの正規母集団 $N(\mu_1,\sigma_1^2)$, $N(\mu_2,\sigma_2^2)$ から，それぞれ大きさ n_1, n_2 の標本を抽出する．それらの不偏分散をそれぞれ U_1^2, U_2^2 とするとき，

$$F = \frac{U_1^2/\sigma_1^2}{U_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1) \qquad \text{が成り立つ．}$$

(証) $\dfrac{(n_1-1)U_1^2}{\sigma_1^2}$, $\dfrac{(n_2-1)U_2^2}{\sigma_2^2}$ はそれぞれ自由度 n_1-1, n_2-1 の χ^2 分布に従うので，比

$$\frac{\frac{(n_1-1)U_1^2}{\sigma_1^2}/(n_1-1)}{\frac{(n_2-1)U_2^2}{\sigma_2^2}/(n_2-1)} = \frac{U_1^2/\sigma_1^2}{U_2^2/\sigma_2^2}$$

は自由度 (n_1-1, n_2-1) の F 分布に従う．

★ 母分散の比の検定 (p.146) で用いられる

Excel の関数

分布	Excel の関数
正規分布	NORM.DIST($\mathbf{x}$, 平均, SD,TF)
正規分布の累積分布関数の逆関数	NORM.INV(確率, 平均, SD)
χ^2 分布	CHISQ.DIST($\mathbf{x}$, 自由度,TF)
左側 $\boldsymbol{t}$ 分布の値	T.DIST($\mathbf{x}$, 自由度,TF)
$\boldsymbol{F}$ 分布の（左側）確率	F.DIST($\mathbf{x}$, 自由度 1, 自由度 2,TF)
χ^2 分布の左側確率の逆関数	CHISQ.INV(確率, 自由度)
$\boldsymbol{t}$ 分布の左側確率の逆関数	T.INV(確率, 自由度)
$\boldsymbol{F}$ 分布の左側確率の逆関数	F.INV(確率, 自由度 1, 自由度 2)

- TF: 関数形式. T (True) は累積, F (False) は密度関数
- 関数の後に .RT を付けると上側 (右側) 確率になる
- 正規分布で NORM.S. とすると標準正規分布になる

まとめ

標本分布

仮説検定や区間推定において重要な役割を演じます．今回学んだのは，正規分布から派生した 3 つの分布です

- χ^2 分布

  分布の「ズレ」を調べるのに使います
- t 分布

  平均の差があるかどうかを調べます
- F 分布

  分散の差があるかどうかを調べます

10 月 16 日の演習問題

問題

次の問いに答えよ。

(1) 確率変数 X が自由度 6 の χ^2 分布に従うとき, $P(Z < a) = 0.05$ となる a の値を求めよ。

(2) 確率変数 T が自由度 13 の t 分布に従うとき, $P(|T| < b) = 0.95$ となる b の値を求めよ。

(3) 確率変数 F が自由度 $(8, 9)$ の F 分布に従うとき, $P(F < c) = 0.05$ となる c の値を求めよ。

次回

「行列」をやります．AIDLE-K にある「行列入門」を読んでおいてください．

課題提出

AIDLE-K にアクセスし，「1016 課題提出」から提出する．提出期限は **10/19 23:59** とする．フィードバック (今日中) も忘れずに．

注意

- ファイル形式に注意し，拡張子をつける (jpg, pdf 等)
- 解答用紙にも名前を書くこと
- **過去のプリントで出していないものがある場合は，超過でもかまわないので出すこと**