データ解析 第9回
t分布・比率の信頼区間

北九州市立大学経済学部

齋藤 朗宏

今日の話題

- t 分布
- 比率の標準誤差，信頼区間

- 実習

t分布

t 分布

- 中心極限定理より，

$$z = \frac{\bar{x} - \mu_x}{\sigma_x / \sqrt{n}}$$

は標準正規分布に従う．一方で，

$$t = \frac{\bar{x} - \mu_x}{s_x / \sqrt{n}}$$

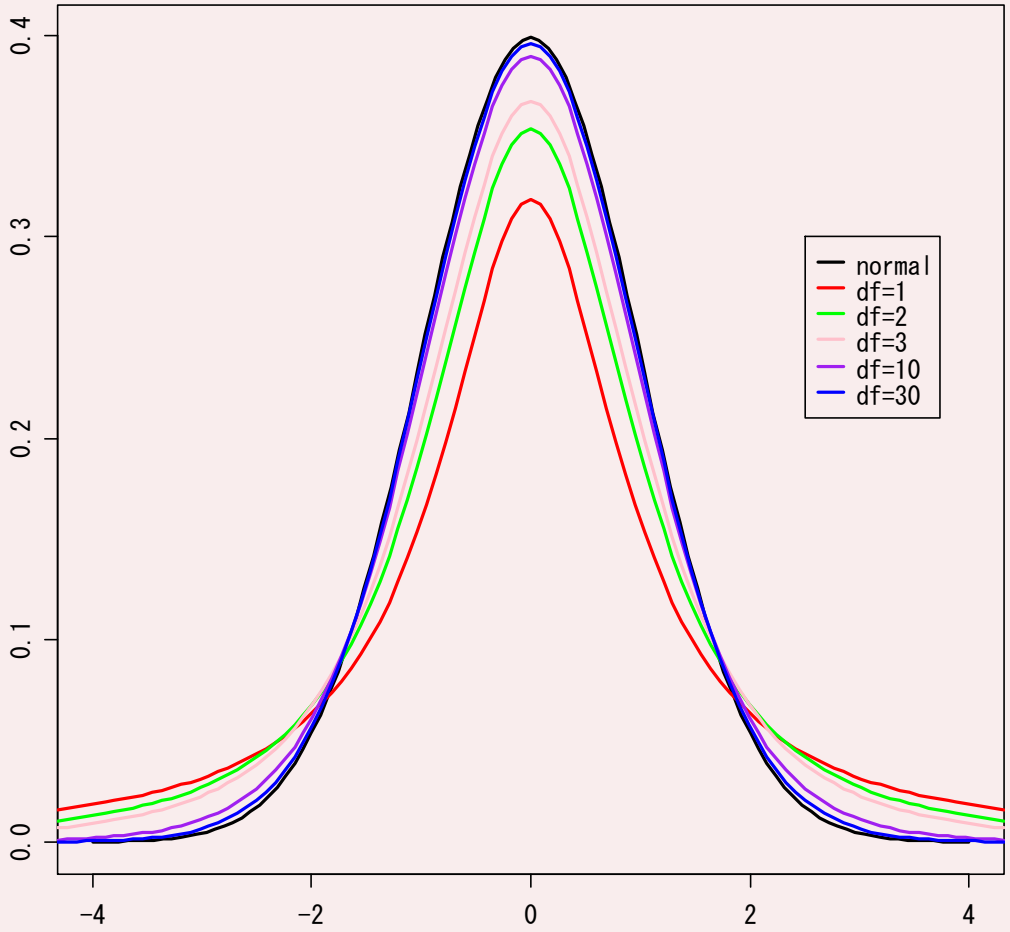
は標準正規分布とは若干異なる分布に従うことが知られている．

- このとき，t が従う分布のことを「t 分布」や，「スチューデントの t 分布」などと呼ぶ．

t 分布

- t 分布は，標本数によって分布の形が決まる．その形は標準正規分布と類似しているが，標本数が少ないときにはやや異なる．
- 標本数が n 個のときには「自由度（Degree of Freedom, df）$n-1$ の t 分布」と呼ぶ．
- $n=30$ を超えると標準正規分布と t 分布の形は概ね一致する．

標準正規分布と t 分布

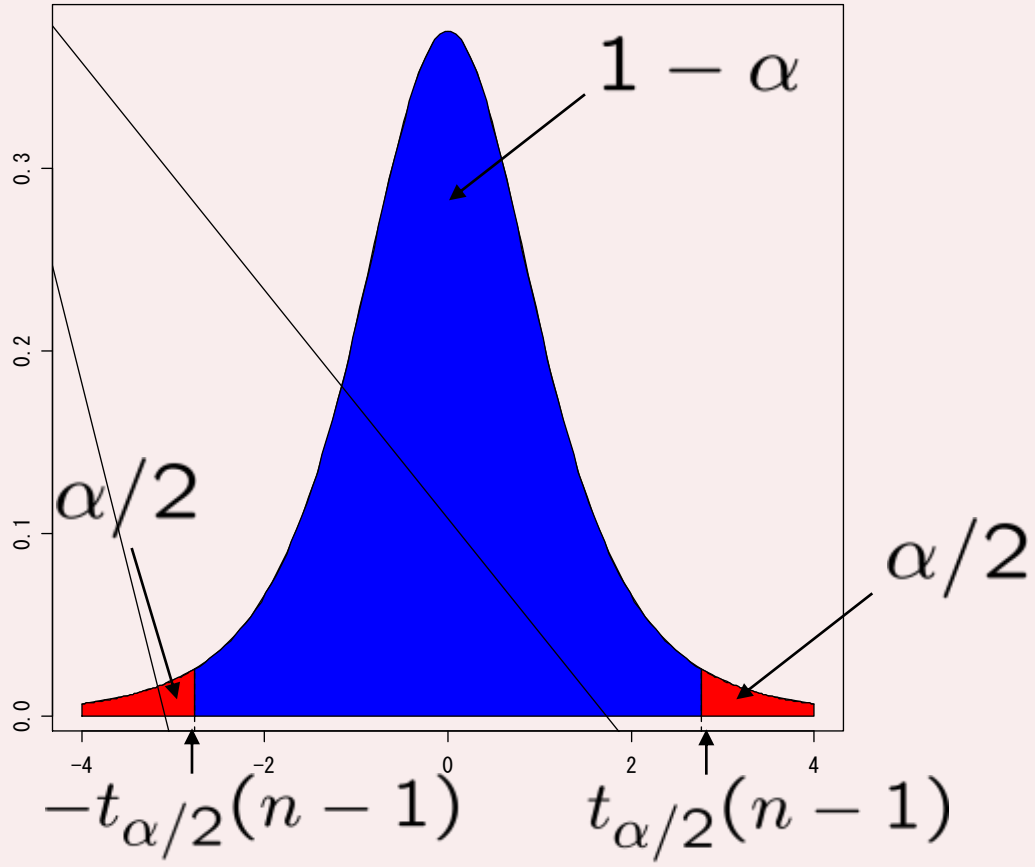


$$P(-1.96 \leq t \leq 1.96) = 0.95$$

にはならない点，自由度ごとに形が違う点に注意．

t 分布（記号の定義）

- 自由度$n-1$の t 分布における確率$1-\alpha$ の範囲を考える．このときの値を$t_{\alpha/2}(n-1)$と表記することとする．



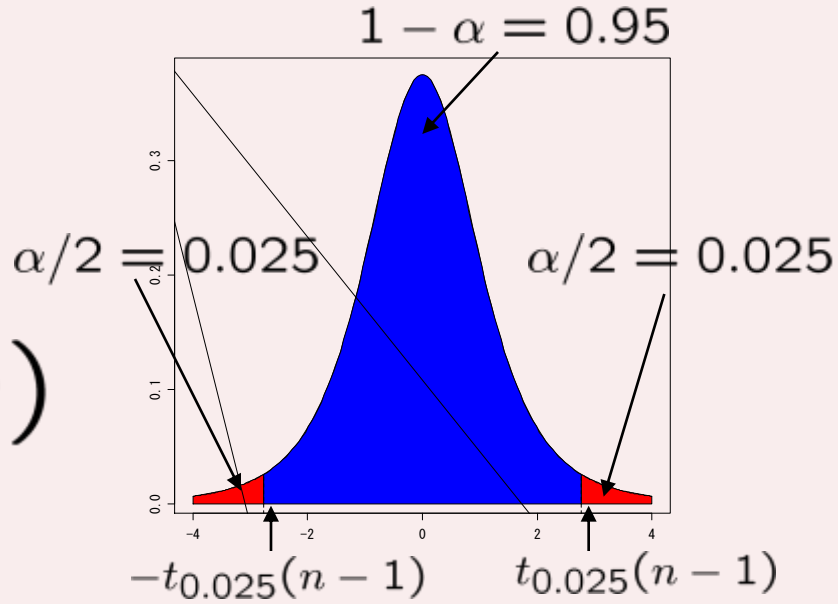
$t_{\alpha/2}(n-1)$ の算出

- 自由度９の t 分布の95%範囲は，$\alpha = 1 - 0.95$ なので，$t_{0.025}(9)$と表記される．
- 簡単なものであれば t 分布表を見るのもよい．
- Excelでも求められる．自由度９の t 分布で95%信頼区間を求めたいのであれば以下の通り．

＝T.INV(0.975,9)

＝T.INV.2T(0.05,9)

どちらも結果は同じ



t 分布における信頼区間

$$P(t_{0.025}(n-1) \leq t \leq t_{0.025}(n-1)) = 0.95$$

$$t = \frac{\bar{x} - \mu_x}{s_x/\sqrt{n}}$$

$$P\left(t_{0.025}(n-1) \leq \frac{\bar{x} - \mu_x}{s_x/\sqrt{n}} \leq t_{0.025}(n-1)\right) = 0.95$$

$$P\left(\bar{x} - t_{0.025}(n-1)\frac{s_x}{\sqrt{n}} \leq \mu_x \leq \bar{x} + t_{0.025}(n-1)\frac{s_x}{\sqrt{n}}\right) = 0.95$$

t 分布を用いた信頼区間

- 一般的に，$1-\alpha$ 信頼区間は上側の式のように与えられる．95%信頼区間は下側の式の通り．

$$\bar{x} \pm t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}$$

$$\bar{x} \pm t_{0.025}(n-1)\frac{s}{\sqrt{n}}$$

- 正規分布を用いた場合の95%信頼区間は以下の通りであった．

$$\bar{x} \pm 1.96\frac{\sigma_x}{\sqrt{n}}$$

- t分布の場合，その形は自由度ごとに違うので，標準正規分布と違い，毎回計算する必要がある．

t 分布を用いた信頼区間

- 標本平均173，標準偏差8，標本数10のとき，母平均の95%信頼区間は

$$\bar{x} \pm t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}$$

$$173 \pm t_{0.025}(9)\frac{8}{\sqrt{10}}$$

$$173 \pm 2.262\frac{8}{\sqrt{10}}$$

$$173 \pm 5.723$$

- となり，信頼区間は167.28から178.72．この区間は母標準偏差を用いた場合より広い．

比率の標準誤差，信頼区間

比率の標準誤差

- 比率を，「失敗のとき0，成功のとき1という値を取る確率変数の平均値」と考える．
  - フリースローの成功率を考える，10投中8投成功したならば，$(1 \times 8 + 0 \times 2)/10$で平均は求められ，平均値＝比率は0.8となる．
- このことを利用すると，比率の標準偏差は以下の通りとなる．

$$\sigma_x = \sqrt{P(1-P)}$$

比率の標準誤差

- 中心極限定理を利用すると，標準誤差は次の通りとなる.

$$\sigma_{\bar{x}} = \sqrt{\frac{P(1-P)}{n}}$$

- つまり，比率の標準誤差は確率と標本数に依存し，その他の値には依存しない.

視聴率調査

- ビデオリサーチ社では，全国27地域において視聴率調査を行っている．特に関東地区，関西地区，名古屋地区の3地区で大規模な調査は行われていて，それぞれ600世帯が調査対象になっている．
- このことと，視聴率がわかれば視聴率の信頼区間も算出できる．

視聴率調査

- ドラマ“相棒”の2012年11月21日における関東地区の視聴率は16.4%であった．このドラマの視聴率の95%信頼区間は？

調査対象は600世帯なので，標準誤差は

$$\sigma_{\bar{x}} = \sqrt{\frac{0.164(1-0.164)}{600}} = 0.015$$

標本数600であれば，t分布はほぼ正規分布と見なせる．よって，

$0.164 \pm 1.96 \times 0.015$ となり，

13.5%から19.3%となる．

実習１

- “グループ2”から，身長の標本平均，標準偏差を求め，そこから標準誤差を推定せよ.
- 母平均の95%信頼区間を求めよ.

実習２

- 11月3日放映の“プロ野球日本シリーズ第6戦”の関東地区における視聴率は23.3%であった．調査対象を600世帯としたとき，視聴率の95%信頼区間を求めよ．ただし，視聴率の推定値の分布をｔ分布と見なした場合と正規分布と見なした場合での信頼区間を両方求めよ．
- 11月19日放映の“ペイルライダー”の関東地区における視聴率は2.7%であった．調査対象を600世帯としたとき，視聴率の95%信頼区間を求めよ．ただし，視聴率の推定値は正規分布に従うとする．
- 上3つの信頼区間から何がわかるか確認せよ．

補足資料

比率の分散，標準誤差

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

[平均値は比率そのものなので]

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

[x_iは1か0いずれか．場合分けする]

$$= \frac{1}{N}\left(NP(1-P)^2 + N(1-P)(0-P)^2\right)$$

1であるデータの個数

0であるデータの個数

比率の分散，標準誤差

$$
\begin{aligned}
&= \left(P(1-P)^2 + (1-P)P^2\right) \\
&= P(1-P)\,(1-P+P) \\
&= P(1-P)
\end{aligned}
$$

$$
\begin{aligned}
\sigma &= \sqrt{P(1-P)} \\
\frac{\sigma}{\sqrt{n}} &= \sqrt{\frac{P(1-P)}{n}}
\end{aligned}
$$

比率の標準誤差