統計学ノートNo. 7
推　定

担当教員：蛭川雅之

研究室：紫英館3階345研究室

メール： hirukawa@econ.ryukoku.ac.jp

オフィスアワー：毎週金曜日3講時

1　点推定

推測統計の中身

- **推測統計は推定と検定を含む。**
  1. **推定：標本から母集団分布の母数（例：平均、標準偏差…）の値を推測する。**
  2. **検定：分析者が想定する母集団に関する仮説が正しいかを標本に照らしあわせて客観的に判断する。**
- **推定は以下の2種類に大別される。**
  1. **点推定：母数を唯一の値で推定する。**
  2. **区間推定：母数をある区間で推定する。**

点推定：一例

- **母平均μを大きさnの標本$X_1, X_2, \ldots, X_n$を用いて推定したい。**
- **既に学んだ通り、標本平均**

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

が候補の一つとなる。

  - **このとき、標本平均$\bar{X}$を母平均μの推定量であるという。**

幾つかの疑問

1. **推定量と推定値の違いは何か？**

- **推定量：**
  - 推定の方法・構造を指す。
  - 標本の関数であり、確率変数である。
- **推定値：**
  - 推定量から算出された数値を指す。

幾つかの疑問（つづき）

2. **未知母数（例えば、母平均μ）の推定量はただ一つか？**

   - **一つに限定されない。**
   - **不偏性・一致性を持つ推定量は幾つも存在する。**
     - **例：モーメント推定量、最尤推定量…**

推定量の望ましい性質

- **理想的には、以下の性質を同時に満たす推定量が望ましい。**
  1. **不偏性**
  2. **一致性**
  3. **有効性**
- **実際には、状況に応じていずれかの基準を優先する。**
- **以下、推定したい未知母数をθ、その推定量を$\hat{\theta}$と表記する。**

不偏性

- **推定量$\hat{\theta}$が**

$$E(\hat{\theta}) = \theta$$

を満たすとき、$\hat{\theta}$はθの不偏推定量であるという。

  - **$\hat{\theta}$は「不偏性を持つ」ともいう。**

- **$\hat{\theta}$がθの不偏推定量であるとき、$\hat{\theta}$のバイアス**

$$b(\hat{\theta}) = E(\hat{\theta}) - \theta$$

はゼロである。

一致性

- **推定量$\hat{\theta}$が$n \to \infty$に対し**

$$\hat{\theta} \xrightarrow{p} \theta$$

を満たすとき、$\hat{\theta}$はθの一致推定量であるという。

- **$E(\hat{\theta}) = \theta$かつ$Var(\hat{\theta}) = 0$が理想であるが、現実には成立しそうにない。**
  - **十分大きなnに対し、これらが近似的に成立することは依然として望ましい。**

一致性（つづき）

- **一致性は推定量が最小限満たすべき性質である。**
  - **推定量が一致性を持たない。**

    ⇒大標本を利用して計算する意味はあるか？

有効性

- **θに関して不偏性を持つ2つの推定量$\hat{\theta}$と$\hat{\theta}^*$に対し**

$$Var(\hat{\theta}) \leq Var(\hat{\theta}^*)$$

が成り立つとき、$\hat{\theta}$は$\hat{\theta}^*$より有効であるという。

有効性（つづき）

- **θに関して不偏性を持つ推定量$\hat{\theta}$がどのような不偏推定量$\hat{\theta}^*$に対しても**

$$Var(\hat{\theta}) \leq Var(\hat{\theta}^*)$$

を満たすとき、 $\hat{\theta}$を最小分散不偏推定量あるいは有効推定量であるという。

2　区間推定

区間推定とは…

- **母数がどのような値をとるかを一点で求めるのでなく、特定の確実さでどのような区間に入っているかを求める。**
  - **特定の確実さ ＝ 信頼係数**

    （例： $100(1-\alpha)\% = 90\%, 95\%, 99\%$ **）**
  - **区間 ＝ 信頼区間**
  - **信頼係数↑(↓)**

    ⇒ 誤っている「確率」↓(↑)・区間幅↑(↓)

区間推定とは…（つづき）

- **最もよく使われる「母集団の平均に関する区間推定」に特化する。**

1. **正規母集団：**
   - **正規分布（分散既知の場合）**
   - **t分布（分散未知の場合）**
2. **未知の母集団：**
   - **正規分布（中心極限定理に基づく正規近似）**

区間推定とは…（つづき）

- **同一の信頼係数を持つ区間の作り方は無数にある。**
  - **区間幅を最短にする観点から、信頼区間を左右対称にとる。**

3　分散既知の正規母集団の母平均に関する区間推定

正規母集団に関する区間推定

- **まず、母集団が正規分布であると仮定して話を進める。**
  - **目標：母平均の区間推定**
- **次の2つの場合について母平均の区間推定を考える。**
  1. **母分散が既知の状態で母平均を推定する。**
  2. **母分散が未知の状態で母平均を推定する。**
- **ケース1には正規分布、ケース2にはt分布を利用する。**

母分散が既知の場合：例題

コンビニのおにぎりを自動的に生産する機械がある。この機械はおにぎりの重さをいろいろ調節することができるが、重さには誤差が生じる。

この機械で製造されるおにぎりの重さ全体が正規母集団に従っており、母分散が100 （即ち、標準偏差が10g）であることが知られているとする。この機械でおにぎり25個を作ってみたとき、その重さの平均は80gであった。

製造されるおにぎりの重さの平均に関する95%信頼区間を求めよ。

母分散が既知の場合：例題（つづき）

- **問題文から次の内容を読み取ることができる。**

1. **おにぎりの重さは平均μ（未知）、分散$\sigma^2 = 100$（既知）の正規分布に従う。**
   - **ケース1に関する例題である点を確認せよ。**
2. **データ数$n = 25$の標本に関する標本平均は$\bar{X} = 80$である。**

母分散が既知の場合：例題（つづき）

- **既に学んだ通り、標本平均$\bar{X}$の分布は**

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

$$\left(= N\left(\mu, \frac{100}{25}\right) = N(\mu, 4)\right)$$

である。

母分散が既知の場合：例題（つづき）

- **標準化により、**

$$\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim N(0,1)$$

である。

- **ここで、信頼係数を95%にとると、**

$$0.95 = Pr\left\{-1.96 \leq \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \leq 1.96\right\}$$

が成り立つ。

母分散が既知の場合：例題（つづき）

- **さらに、**

$$-1.96 \leq \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \leq 1.96$$

をμについて解くと、

$$\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}} \cdots (1)$$

が得られる。

母分散が既知の場合：例題（つづき）

- (1)**の左辺と右辺に**

$$\bar{X} = 80, \frac{\sigma}{\sqrt{n}} = \sqrt{4} = 2$$

を代入することにより、母平均μ**に関する**95%**信頼区間**

$$80 - 1.96 \times 2 \leq \mu \leq 80 + 1.96 \times 2$$
$$\Rightarrow 76.08 \leq \mu \leq 83.92$$

を得る。

4　分散未知の正規母集団の母平均に関する区間推定

母分散が未知だとすると…

- **母分散σ^2が未知の正規母集団について、母平均μの区間推定をどのように行えばよいか？**
  - **この区間推定を実行するには、母平均μを除き、標本を使って計算できる統計量に依存する分布を考える必要がある。**

t統計量の導入

- **$N(\mu, \sigma^2)$から大きさnの標本$X_1, X_2, \ldots, X_n$を取り出す。**

1. **標本平均**

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

を計算する。

t統計量の導入（つづき）

2. **標本標準偏差**

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

を計算する。

3. **次のようにt統計量を定義する。**

$$t = \frac{\sqrt{n-1}(\bar{X} - \mu)}{\hat{\sigma}}$$

t統計量の分布

- **$N(\mu, \sigma^2)$から取り出された大きさnの標本$X_1, X_2, \ldots, X_n$に対して定義される統計量**

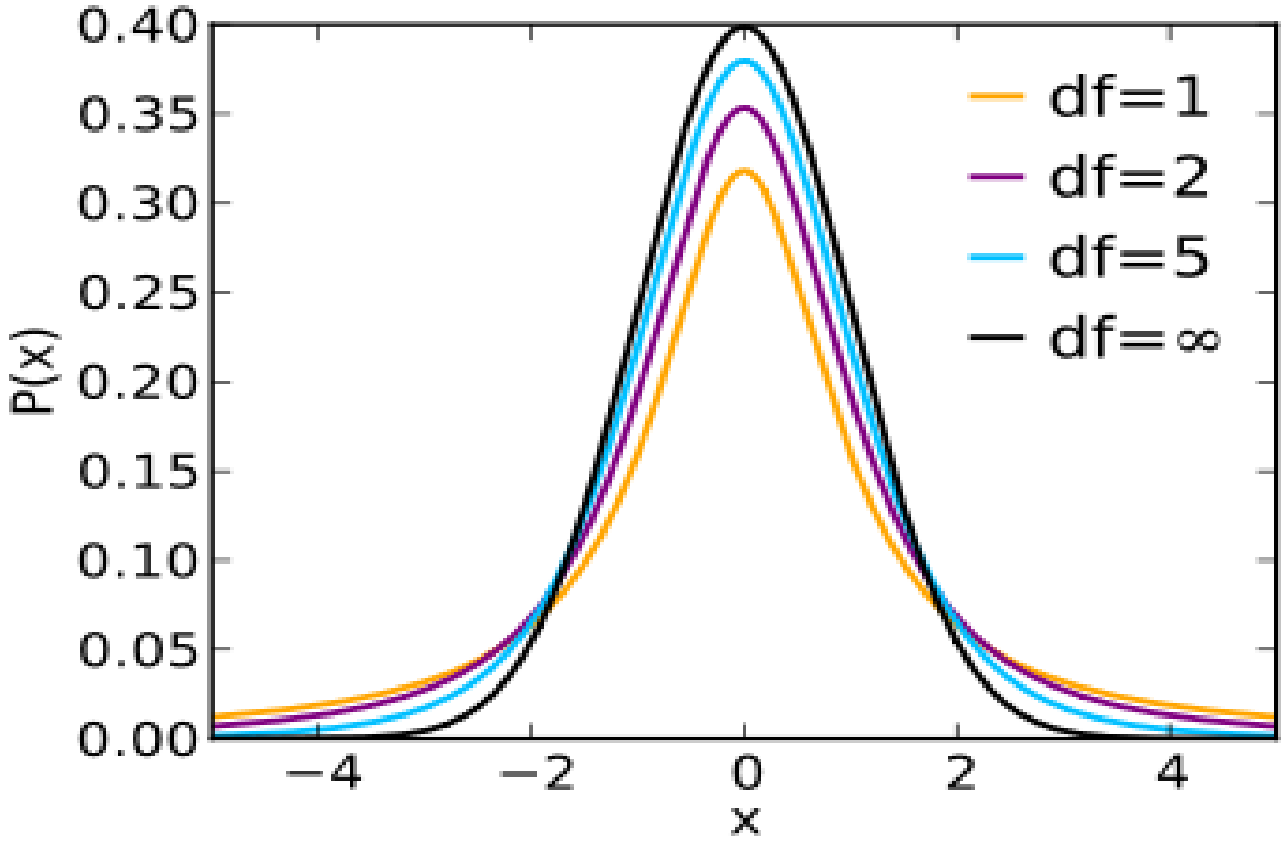
$$t = \frac{\sqrt{n-1}(\bar{X} - \mu)}{\hat{\sigma}}$$

は自由度$(n-1)$のt分布に従う。

  - **このことを以下のように表記する。**

$$t = \frac{\sqrt{n-1}(\bar{X} - \mu)}{\hat{\sigma}} \sim t(n-1)$$

t統計量の分布（つづき）



（重要）自由度∞のt分布は標準正規分布である。

t分布表

- **t分布表は所与の自由度に対するt分布の$100q\%$分位点を与える。**
  - **Excelの関数`tinv`を利用してもよい。**
- **t分布表を利用する際、t分布が正規分布と同様左右対称である点に注意する。**

t分布による区間推定：例題

ある蝶の体長の母集団は正規分布であることが知られている。観測した６個体の体長がそれぞれ76mm、85mm、82mm、83mm、76mm、78mmであった場合、母平均μに関する95%信頼区間を求めよ。

t分布による区間推定：例題（つづき）

- **問題文から次の内容を読み取ることができる。**

1. **蝶の体長は平均μ、分散σ^2がともに未知の正規分布に従う。**
   - **ケース2に関する例題である点を確認せよ。**
2. **データ数$n = 6$の標本**

$$X_1 = 76, X_2 = 85, X_3 = 82$$
$$X_4 = 83, X_5 = 76, X_6 = 78$$

を得た。

t分布による区間推定：例題（つづき）

- **まず、**

$$t = \frac{\sqrt{n-1}(\bar{X}-\mu)}{\hat{\sigma}} \sim t(n-1)$$

である。

- **ここで、信頼係数を**95%**にとり、**$t(n-1)$**の上側確率が**2.5%**となる分位点を**$t_{0.025}\left(= t_{0.025}(n-1)\right)$**と表記すると、**

$$0.95 = Pr\left\{-t_{0.025} \leq \frac{\sqrt{n-1}(\bar{X}-\mu)}{\hat{\sigma}} \leq t_{0.025}\right\}$$

が成り立つ。

t分布による区間推定：例題（つづき）

- **さらに、**

$$-t_{0.025} \leq \frac{\sqrt{n-1}(\bar{X}-\mu)}{\hat{\sigma}} \leq t_{0.025}$$

をμについて解くと、

$$\bar{X} - t_{0.025}\frac{\hat{\sigma}}{\sqrt{n-1}} \leq \mu \leq \bar{X} + t_{0.025}\frac{\hat{\sigma}}{\sqrt{n-1}} \cdots (2)$$

が得られる。

t分布による区間推定：例題（つづき）

- **ここで、標本平均と標本分散はそれぞれ**

$$\bar{X} = \frac{76 + 85 + 82 + 83 + 76 + 78}{6} = 80$$

$$\hat{\sigma}^2 = \frac{(-4)^2 + 5^2 + 2^2 + 3^2 + (-4)^2 + (-2)^2}{6} = \frac{37}{3}$$

である。

- **また、t分布表から**

$$t_{0.025} = t_{0.025}(5) = 2.571$$

である。

t分布による区間推定：例題（つづき）

- **これらを(2)の左辺と右辺に代入することにより、母平均μに関する95%信頼区間**

$$80 - 2.571 \times \frac{\sqrt{37/3}}{\sqrt{5}} \leq \mu \leq 80 + 2.571 \times \frac{\sqrt{37/3}}{\sqrt{5}}$$

$$\Rightarrow 75.96 \leq \mu \leq 84.04$$

を得る。

5　中心極限定理

正規母集団の仮定は強すぎる…

- **母集団が正規分布であることを仮定しない（あるいは、仮定することに無理がある）場合、母平均の区間推定をどのように行うか？**

- **実は、母集団の分布形を特定せずに区間推定を行う方法がある。**

1. **中心極限定理による正規近似**
2. **ノンパラメトリック法**

統計量Z_n

- **母集団の分布は平均μ、分散σ^2を持つとする。**
  - **分布形は特定していない点に注意せよ。**
- **この母集団から大きさnの標本$X_1, X_2, \ldots, X_n$を取り出す。**

統計量Z_n（つづき）

- **このとき、標本平均**

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

の分布の平均はμ、分散はσ^2/nである。

統計量Z_n（つづき）

- **統計量Z_nは標本平均$\bar{X}$を標準化したものとして定義される。**

$$Z_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

  - **Z_nの平均は0、分散は1である。**

中心極限定理【Ver. 1】

- **平均μ、分散σ^2を持つ同一の母集団から大きさnの標本$X_1, X_2, \ldots, X_n$を取り出し、その標本平均**

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

を計算する。nが十分大きいとき、統計量

$$Z_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

の分布は標準正規分布に近づく。

中心極限定理（つづき）

1. **中心極限定理は以下のように言い換えることができる。**

   - **平均μ、分散σ^2を持つ母集団からn個のデータを観測し、標本平均を計算する。データ数nが十分大きいとき、この標本平均の分布は$N(\mu, \sigma^2/n)$に近づく。**

   - **独立なn個のデータの和がおおよそ$N(n\mu, n\sigma^2)$に従うと考えてもよい。**

中心極限定理（つづき）

2. **中心極限定理は正規母集団を仮定しなくても成り立つ。**
   - **中心極限定理は平均μ、分散σ^2を持つ任意の母集団に対して成り立つ。**
3. **中心極限定理は「標本平均の分布が正規分布で近似できる」ことを意味する。**
   - **標本平均の正確な分布を明示できなくても、データ数が十分大きければ、標本平均の分布を正規分布と見なして差し支えない。**

中心極限定理（つづき）

4. **「データ数nが十分大きい」とは具体的にはデータ何個のことか？**

   - **正規分布のあてはまりの良さは母集団の分布形に依存する。**
   - **データ数に明確なガイドラインはないが、母集団が二項分布や一様分布などの場合、$n = 20$~30 程度でも十分良い近似が得られる。**

中心極限定理【Ver. 2】

- **中心極限定理【Ver. 1】で定義した統計量**

$$Z_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

について、標準偏差σを標本標準偏差

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

に置き換える。

中心極限定理【Ver. 2】（つづき）

- **このようにして得られる統計量**

$$Z_n^* = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}}$$

の分布も、nが十分大きいとき、標準正規分布に近づく。

6　中心極限定理の重要な応用

例題 1

ある地区のスーパーについて特定の銘柄のインスタント・コーヒーの価格を調べた結果、420円から600円までの一様分布とみなして差し支えないことがわかった。会合のためにこのインスタント・コーヒー20個を用意したいが予算は1万円である。予算内でコーヒーを20個買うことのできる確率はいくらか。

例題１（つづき）

- **問題文から次の内容を読み取ることができる。**

1. **インスタント・コーヒーの価格は一様分布である。**
   - **この分布の平均は510、分散は2,700である。**
   - **母集団は正規分布でないため、中心極限定理を応用する。**
2. **データ数$n = 20$の標本$X_1, X_2, \ldots, X_{20}$を得た。**
3. **以上の情報をもとに、データの和が1万円以下になる確率を知りたい。**

例題１（つづき）

- **この問題を解くためには、平均μ、分散σ^2の母集団から抽出したn個のデータの和**

$$S_n = \sum_{i=1}^{n} X_i$$

を正規分布で近似する必要がある。

  - **中心極限定理から、データ数nが十分大きいとき、データの和S_nの分布は$N(n\mu, n\sigma^2)$で近似できる。**

例題１（つづき）

- **母平均$\mu = 510$、母分散$\sigma^2 = 2{,}700$を利用すると、和$(S_n =)S_{20}$の分布は以下のような平均と分散を持つ正規分布で近似できる。**
  - **平均：** $n\mu = 20 \times 510 = 10{,}200$
  - **分散：** $n\sigma^2 = 20 \times 2{,}700 = 54{,}000$

例題１（つづき）

- **従って、求める確率は**

$$Pr(S_{20} \leq 10{,}000)$$

$$= Pr\left(\frac{S_{20} - 10{,}200}{\sqrt{54{,}000}} \leq \frac{10{,}000 - 10{,}200}{\sqrt{54{,}000}}\right)$$

$$\approx Pr(Z_{20} \leq -0.86)$$

$$\approx \Phi(-0.86)$$

$$= 1 - \Phi(0.86)$$

$$= 1 - 0.8051$$

$$= 0.1949$$

例題２

あるテレビ局のプロデューサーX氏は制作番組の視聴率目標として20%を想定している。一方、600世帯に対する視聴率調査の結果、この番組を視聴していた世帯は90世帯であった。このデータを利用して視聴率に関する95%信頼区間を求めよ。X氏の視聴率目標は達成できたといってよいか。

例題２（つづき）

- **問題文から次の内容を読み取ることができる。**

1. **母集団（＝視聴者全体）内の世帯が番組を視聴したか否かはベルヌーイ分布に従う。**
   - **母集団の視聴率をpとする。**
   - **正規母集団でないため、中心極限定理を応用する。**
2. **データ数$n = 600$の標本$X_1, X_2, \ldots, X_{600}$を得た。**
3. **以上の情報をもとに、視聴率（母平均）に関する95%信頼区間を求めたい。**

例題２（つづき）

- **中心極限定理【**Ver. 2**】により、**

$$Z_n^* = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\hat{\sigma}}$$

に対し信頼係数を95%**にとると、**

$$0.95 \approx Pr\left\{-1.96 \leq \frac{\sqrt{n}(\bar{X} - \mu)}{\hat{\sigma}} \leq 1.96\right\}$$

が成り立つ。

例題２（つづき）

- **さらに、**

$$-1.96 \leq \frac{\sqrt{n}(\bar{X} - \mu)}{\hat{\sigma}} \leq 1.96$$

をμについて解くと、

$$\bar{X} - 1.96\frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\hat{\sigma}}{\sqrt{n}} \cdots (3)$$

が得られる。

例題２（つづき）

- **ベルヌーイ分布に従う確率変数Xの平均は母集団比率（＝成功確率）pである。**
  - **そこで、(3)を**

$$\bar{X} - 1.96\frac{\hat{\sigma}}{\sqrt{n}} \leq p \leq \bar{X} + 1.96\frac{\hat{\sigma}}{\sqrt{n}} \cdots (3')$$

と書き改める。

例題２（つづき）

- **ところで、ベルヌーイ分布から取り出された標本の平均と分散はどのように計算されるか？**
- **調査世帯を次の2グループに分ける。**

1. **視聴した世帯**
   - **世帯数：$n_1 = 90$**
2. **視聴しなかった世帯**
   - **世帯数：$n_0 = n - n_1 = 510$**

例題２（つづき）

- **まず、標本平均は次のようになる。**

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{\overbrace{0 + \cdots + 0}^{n_0} + \overbrace{1 + \cdots + 1}^{n_1}}{n}$$

$$= \frac{n_0 \times 0 + n_1 \times 1}{n} = \frac{n_1}{n} = \frac{90}{600} = 0.15$$

  - **標本平均は標本比率（＝標本全体に対する視聴した世帯の割合）に等しいため、通常以下のように表記する。**

$$\bar{X} = \frac{n_1}{n} := \hat{p}(= 0.15)$$

例題２（つづき）

- **さらに、標本分散は次のようになる。**

$$\hat{\sigma}^2 = \frac{\overbrace{(0-\bar{X})^2 + \cdots + (0-\bar{X})^2}^{n_0} + \overbrace{(1-\bar{X})^2 + \cdots + (1-\bar{X})^2}^{n_1}}{n}$$

$$= \frac{n_0 \times \bar{X}^2 + n_1 \times (1-\bar{X})^2}{n}$$

$$= \left(\frac{n-n_1}{n}\right)\bar{X}^2 + \left(\frac{n_1}{n}\right)(1-\bar{X})^2$$

$$= (1-\hat{p})\hat{p}^2 + \hat{p}(1-\hat{p})^2$$

$$= \hat{p}(1-\hat{p})$$

$$= 0.15 \times 0.85$$

例題２（つづき）

- (3′)の左辺と右辺に$\bar{X} = \hat{p}$および$\hat{\sigma} = \sqrt{\hat{p}(1-\hat{p})}$を代入することにより、**母集団比率に関する95%信頼区間**

$$\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \cdots (3'')$$

が得られる。

例題２（つづき）

- **(3′′)の左辺と右辺に$\hat{p} = 0.15, n = 600$を代入することにより、母集団比率pに関する95%信頼区間は**

$$0.15 - 1.96\sqrt{\frac{0.15 \times 0.85}{600}} \leq p \leq 0.15 + 1.96\sqrt{\frac{0.15 \times 0.85}{600}}$$

$$\Rightarrow 0.121 \leq p \leq 0.179$$

となる。

例題２（つづき）

- **<u>結論</u>：視聴率の95%信頼区間はおよそ12%~18%であり、目標である視聴率20%を達成したとは言い難い。**