統計学 2

10. カイ二乗分布と母分散の推定

矢内　勇生

2018年5月17日

高知工科大学　経済・マネジメント学群

今日の目標

- 不偏分散とは何か理解する
- カイ二乗（χ^2）分布とはどんな分布か理解する
- 母分散の推定方法を身につける

分散 (variance)

- 分散：データのばらつき（散らばり具合）を表す統計量
- 0以上の値をとる（データの値がすべて等しいとき0）
- データのばらつきが大きいほど、分散も大きくなる

分散の求め方（復習1）

- 標本分散を表す記号：s^2（σ^2は母分散）

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

標本分散の求め方（復習2）

1. 標本平均を求める

2. 偏差（=観測値 - 標本平均）を各観測値について求める

3. それぞれの偏差を2乗する

4. 偏差の2乗をすべて足して、標本サイズ（n）で割る（偏差^{2}の平均値を求める）

標本分散の求め方（復習3）

- 右のデータの分散を求めてみる

$$s^2 = \frac{16+1+0+4+9}{5}$$
$$= \frac{30}{5} = 6$$

	x	偏差	偏差^2
	1	-4	16
	4	-1	1
	5	0	0
	7	2	4
	8	3	9
計	25	0	30

標本分散の偏り

- 標本をいくつも抽出し、それぞれの標本について標本分散を求める：

標本分散の平均値 < 母分散

➡標本分散を母分散の推定に使うと、（小さい方に）偏ってしまう

★偏りがない分散の求め方は？

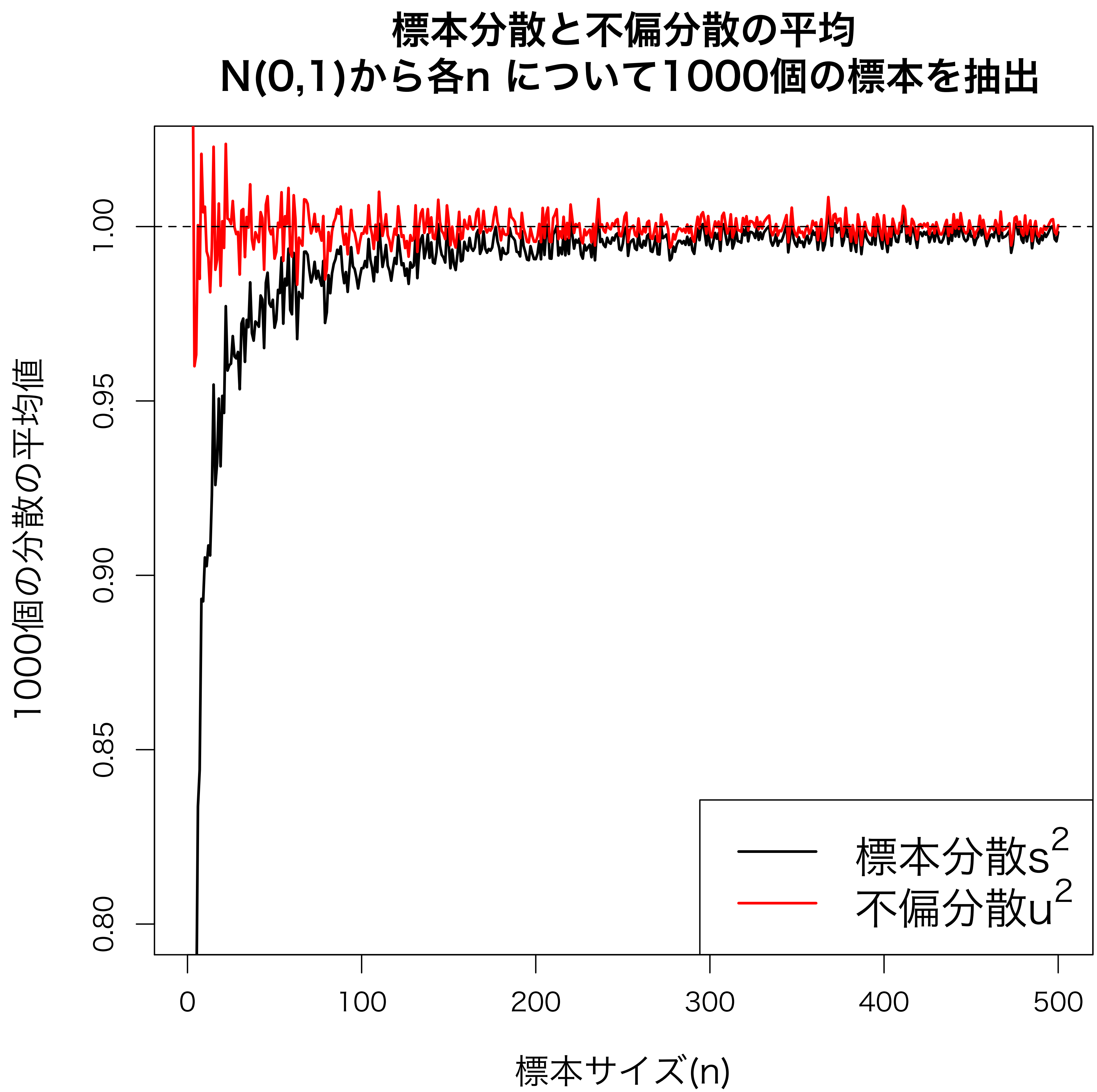
不偏分散（1）（unbiased variance）

- 母分散の偏りのない推定値として使える
- 記号：u^2 で表す
- 分散を求める式の分母を少しだけ小さくする

$$u^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

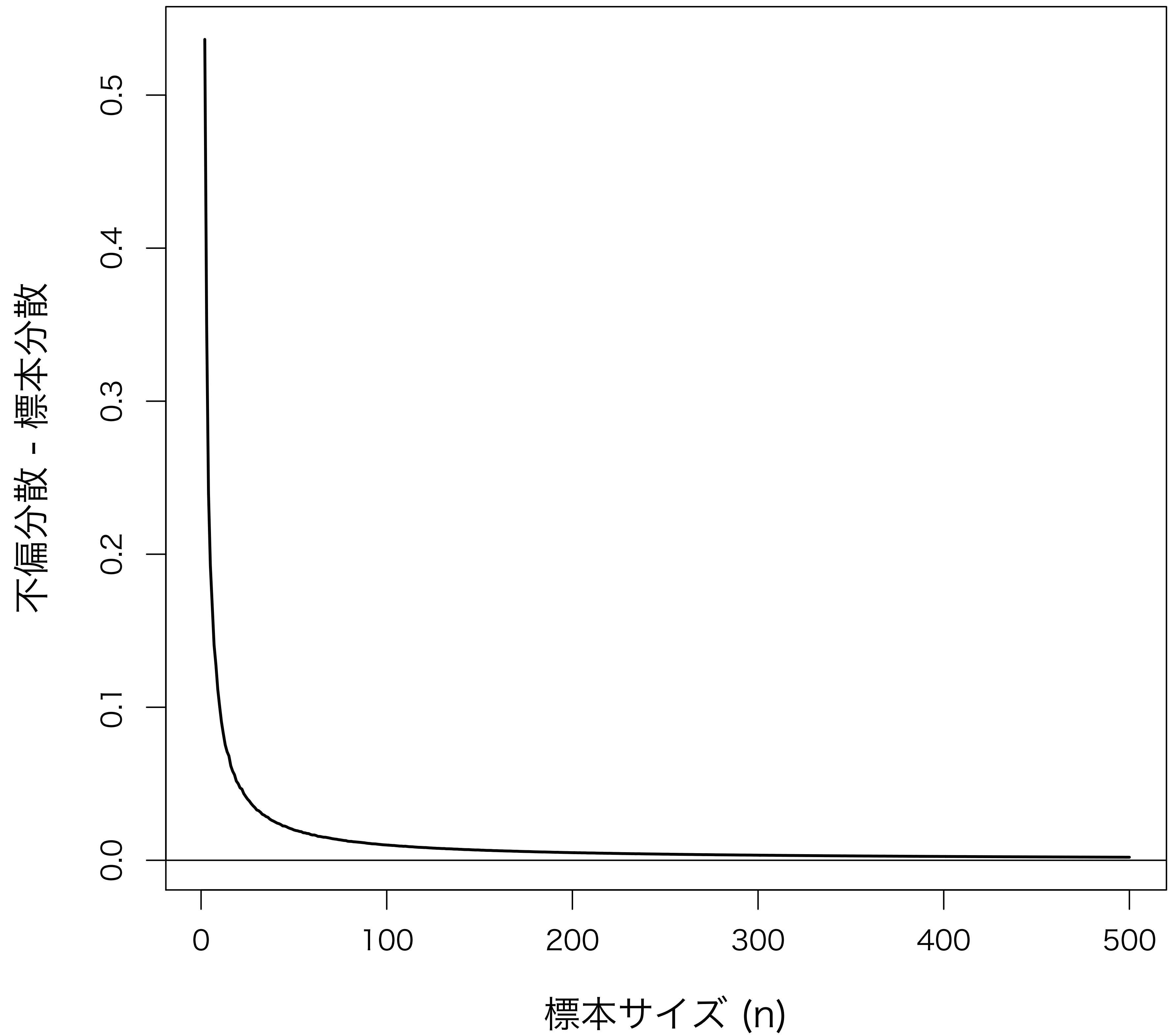
不偏分散（2）

- 不偏分散：母分散の点推定値
- 標準偏差の点推定値は：u（不偏分散の平方根）
    - n が十分大きいときは、標本分散で代用してよい
    - 母平均を知っているときは、標本平均の代わりに母平均を使い、標本分散を求めればよい



標本分散と不偏分散の平均
N(0,1)から各n について1000個の標本を抽出
1000個の分散の平均値
1.00
0.95
0.90
0.85
0.80
0
100
200
300
400
500
標本サイズ(n)
標本分散s2
不偏分散u2

不偏分散と標本分散の差



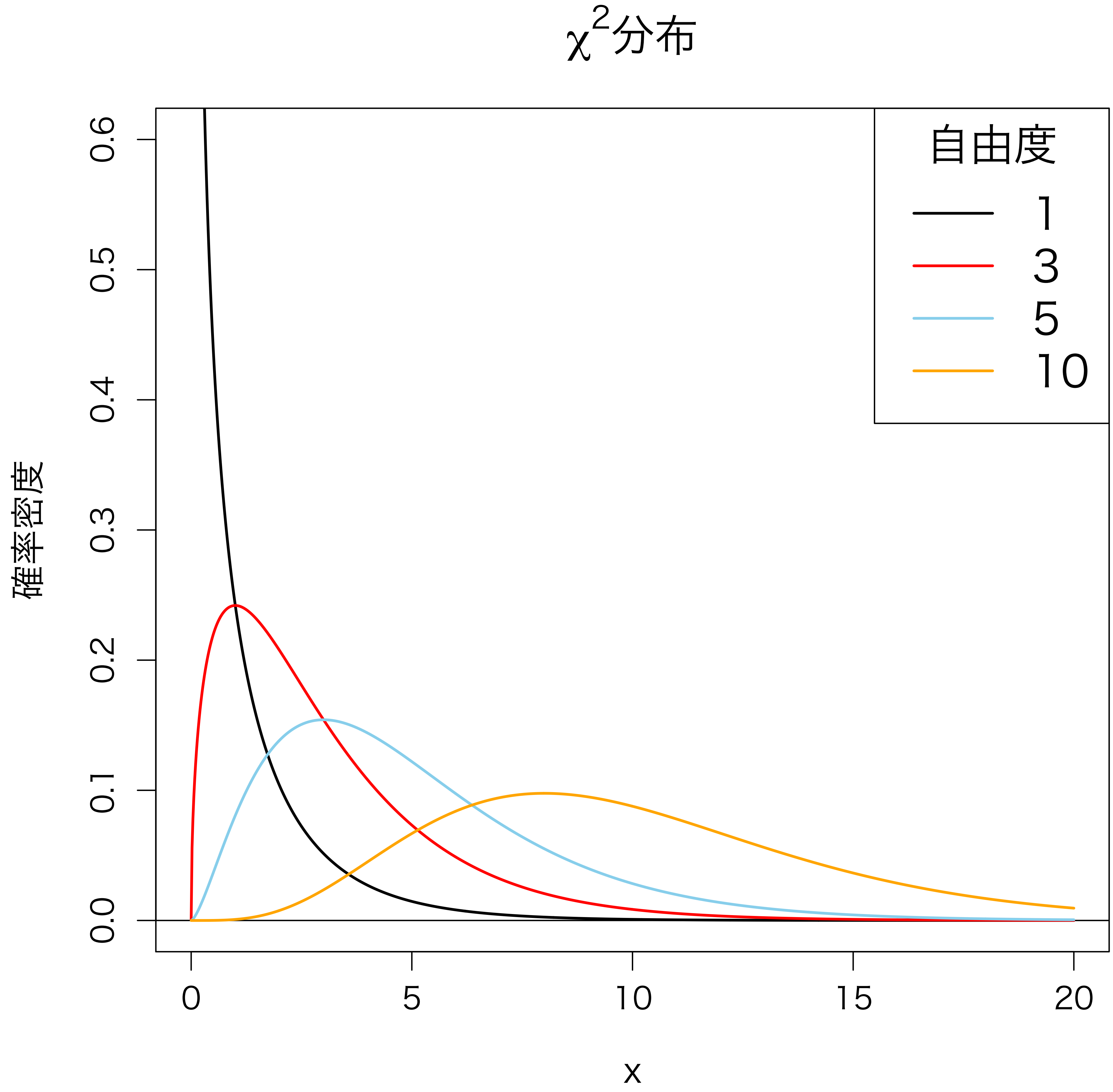
χ^2（カイ二乗）分布 (chi-squared distribution)

- 確率密度関数 *f(x)* は、

$$f(x;k) = \frac{\left(\frac{1}{2}\right)^{\frac{k}{2}}}{\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$$ （x > 0のとき）

$$f(x;k) = 0$$ （x ≤ 0のとき）

ただし、*k* は*x*の自由度、Γ (.) はガンマ関数。


χ2分布
自由度
1
3
5
10
0.0
0.1
0.2
0.3
0.4
0.5
0.6
確率密度
0
5
10
15
20
x

自由度
（degree of freedom: df）

- 自由（独立）に動かせる値の数
- 各統計量に対して自由度が定められる
  - 例：サイズn の標本の場合
    - 標本平均の自由度は n
    - 標本（不偏）分散の自由度は n-1

χ^2分布の定義（μが既知のとき）

- $Z_1, Z_2, \ldots, Z_n$ を（独立な）標準正規分布に従う確率変数とする
  - ただし、 $$z_i = \frac{x_i - \bar{x}}{\sigma}$$
- このとき

$$\chi^2 = z_1^2 + z_2^2 + \cdots z_n^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sigma^2}$$

は自由度 n のカイ二乗分布に従う

χ^2分布の定義（μが未知のとき）

- $Z_1, Z_2, \ldots, Z_n$ を（独立な）標準正規分布に従う確率変数とする
  - ただし、 $$z_i = \frac{x_i - \bar{x}}{\sigma}$$
- このとき

$$\chi^2 = z_1^2 + z_2^2 + \cdots z_n^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sigma^2}$$

は自由度 *n-1* のカイ二乗分布に従う

χ^2分布と不偏分散の関係

$$
\begin{aligned}
\chi^2 &= \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sigma^2} = \sum_{i=1}^{n} \frac{n-1}{n-1} \frac{(x_i - \bar{x})^2}{\sigma^2} \\
&= \sum_{i=1}^{n} \frac{n-1}{\sigma^2} \frac{(x_i - \bar{x})^2}{n-1} = \frac{n-1}{\sigma^2} \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} \\
&= \frac{n-1}{\sigma^2} u^2
\end{aligned}
$$

* 標本分散を使う場合についても同様に考えられる。

不偏分散（u2）の分布

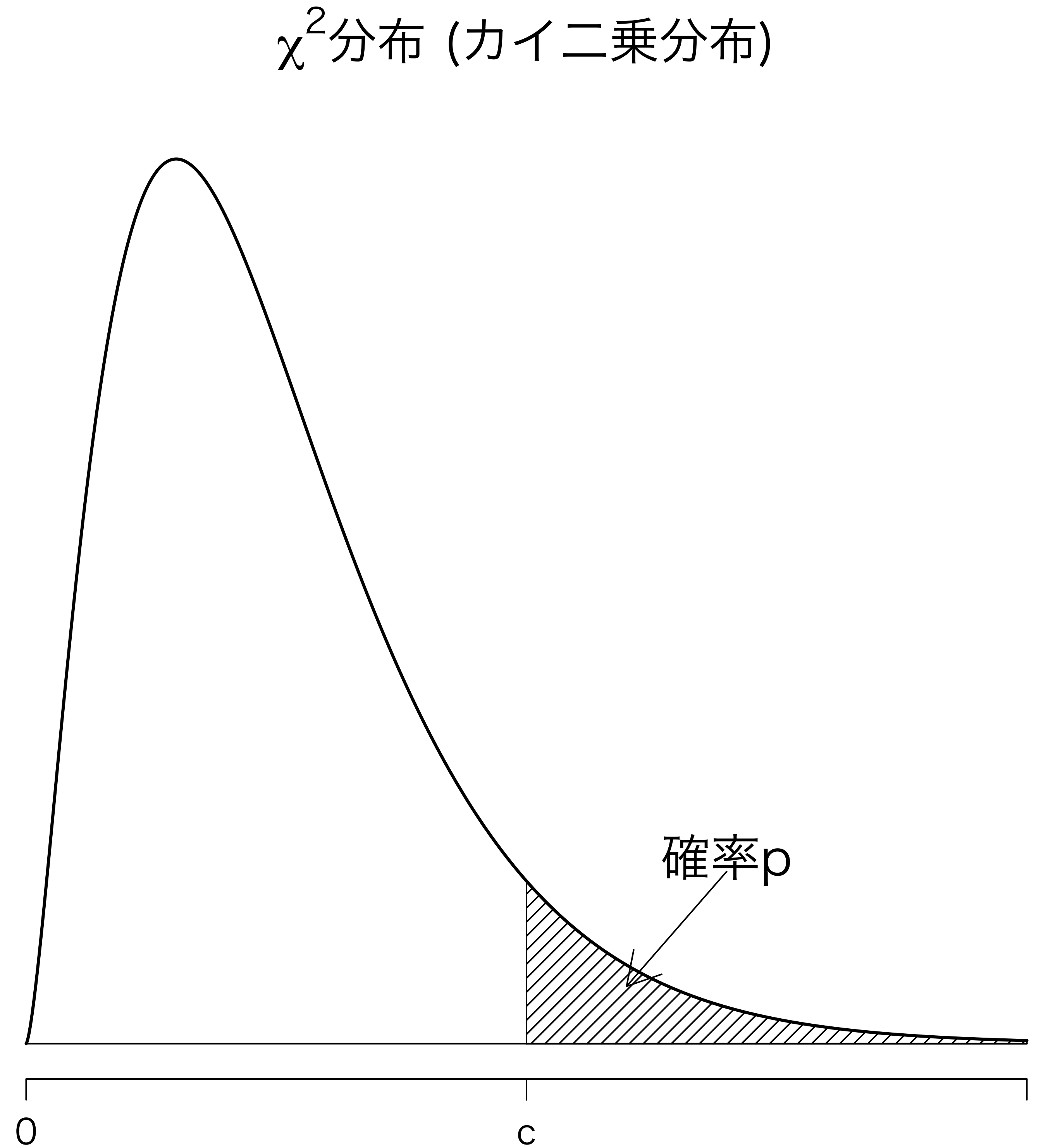
- 母平均を知らないとき、標本サイズ n の標本から求められる不偏分散を u^2 とすると、

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

は自由度 $n-1$ のカイ二乗分布に従う

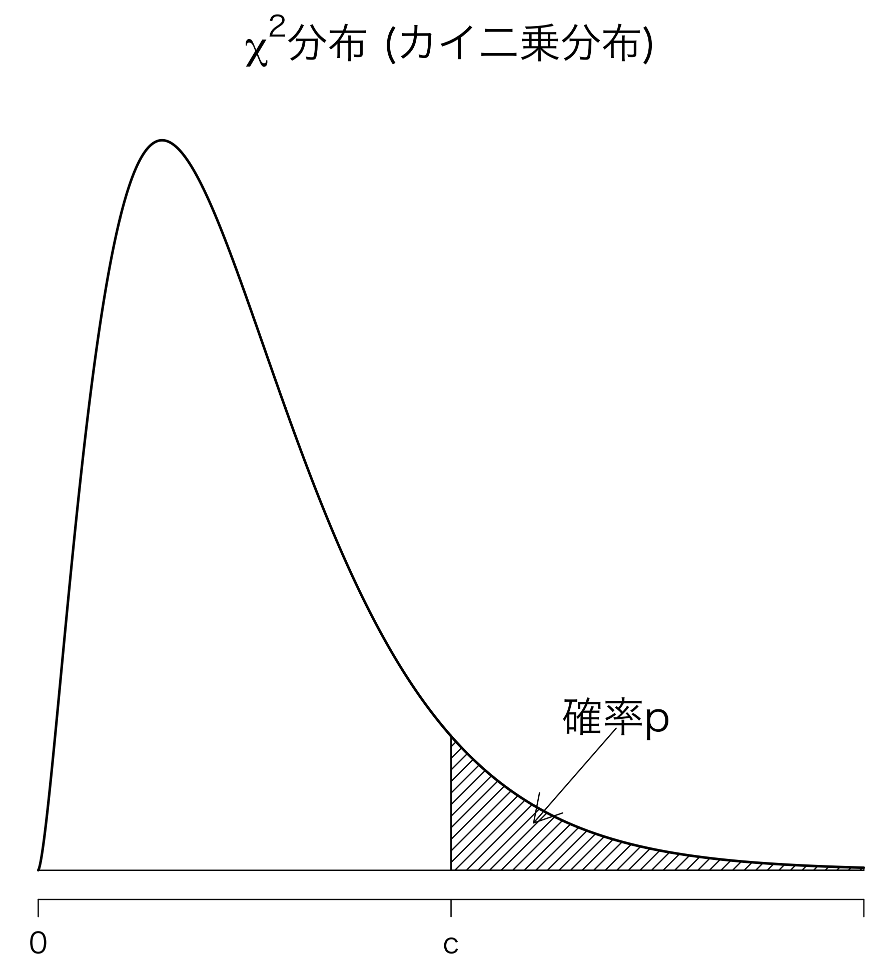
χ^2分布の使い方

- 自由度 (df) を選ぶ
- 選んだ自由度で、自分が知りたい確率 *p*（右図の斜線部）を選ぶ
- *c*を求める : *c* 以上の値をとる確率が*p*



qchisq(p, df) で c を求める

- qchisq(p, df, lower.tail = FALSE)
  - *c* の値が得られる
  - *c* 以下になる確率を設定するときは、lower.tail = TRUE にする



χ^2分布を利用した推定（1）

- 自由度は 5 であるとする
- カイ二乗分布の中央部分95%：カイ二乗分布の図で、左右両端から2.5%分ずつ取り除けばよい

  - 左端2.5%：

    qchisq(p = 0.025, df = 5, lower.tail = TRUE) → 0.831

  - 右端2.5%：

    qchisq(p = 0.025, df = 5, lower.tail = FALSE) → 12.833

➡自由度5のカイ二乗分布の中央部分95%：[0.831, 12.833]

χ^2分布を利用した推定（2）

- n =6のとき（自由度が5のとき）、不偏分散の95%は、

$$0.831 \leq \frac{n-1}{\sigma^2}u^2 \leq 12.833$$

を満たす。

➡ 母分散が σ^2 の正規分布から標本サイズ6の標本を抽出するとき、不偏分散の95%については

$$\frac{0.831}{5}\sigma^2 \leq u^2 \leq \frac{12.833}{5}\sigma^2$$

母分散の95%信頼区間

- n とu^2 は知っている（標本から計算できる）
- σ^2 について不等式を解けばよい

$$0.831 \leq \frac{6-1}{\sigma^2}u^2 \leq 12.833$$

$$0.831\sigma^2 \leq 5u^2 \leq 12.833\sigma^2$$

$$\frac{5}{12.833}u^2 \leq \sigma^2 \leq \frac{5}{0.831}u^2$$

一般的な場合

カイ二乗分布：

$$\chi_L^2 \leq \frac{n-1}{\sigma^2}u^2 \leq \chi_U^2$$

信頼区間：

$$\frac{n-1}{\chi_U^2}u^2 \leq \sigma^2 \leq \frac{n-1}{\chi_L^2}u^2$$

今日のまとめ

- 標本分散は偏っているので不偏分散を使う
  - nが十分大きければ、標本分散を使ってよい
- カイ二乗分布は自由度によって形が決まる
- 不偏分散（標本分散）を少しだけ変形すると、カイ二乗分布に従う
  - μ が既知：自由度 n のカイ二乗分布
  - μ が未知：自由度 $n-1$ のカイ二乗分布