確率統計と情報処理・演習（2007 年度後期）

確率分布

2007 年 12 月 14 日

日本女子大学理学部数物科学科　　今野 良彦

November 30, 2007

今日の講義の目的と概要

- 推定について
  - 母集団と標本
  - 母数（母平均・母分散）
  - 推定量と推定値
  - 点推定法：推定量の性質について
    - 不偏推定量
    - 有効推定量：最良線形不偏推定量
  - 区間推定法
    - 信頼係数と信頼区間（ここからは来週）
  - 正規母集団に対する母平均の区間推定
    - 母分散が既知の場合の母平均の区間推定
    - 母分散が未知の場合の母平均の区間推定

母集団と標本

⋆ 母集団：ある集団の特徴を調査したいとき，調べてたい集団の全体

⋆ 例：

- 2006 年度に日本全体で小学校に入学したときの生徒の身長の平均が知りたい ⟶ 2006 年度に日本全体で小学校に入学したときの生徒全員が母集団

- ある工場で生産されている電球の寿命が知りたい！たとえば，1000 時間以上の寿命があるかどうか？ ⟶ 母集団は母集団は生産される電球すべて！

⋆ 全数調査：母集団の全員を調査すること

⋆ 標本調査：母集団から抜き出された一部分だけを調査すること

⋆ 標本：母集団から抜き出された一部分のこと

推定とは

⋆　母集団のすべてを調査しない！

⋆　母集団の個体の特性値の分布を確率分布でモデル化（まねる）する．この確率分布のことを**母集団分布**という！

⋆注意： 分布とは，個体の特性値がどのあたりにどのくらいあるかを表現するものである！

⋆ 目標：母集団分布を特定したい！できなければ，母集団分布の平均や分散を推測したい！

⋆ 母数：母集団を特定するために役にたつ指標のことをいう．例としては，母集団分布の平均と分散．母集団分布の確率密度関数も母数の例となる．特にことわらないかぎり，推定の対象となる母数は 1 次元として議論を進めていく．

⋆ 母平均：母集団分布の平均

⋆ 母分散：母集団分布の分散

⋆ 仮定：個体の特性値は「連続量」で，母集団分布は正規分布と仮定する．

この仮定より

母集団分布の母平均と母分散がわかる $\Longleftrightarrow$ 母集団分布がわかる！

なぜならば，正規分布の確率密度関数は

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

で

$$\text{母平均：}\quad \mu = \int_{-\infty}^{\infty} x f(x)\,dx$$

$$\text{母分散：}\quad \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx$$

推定法について

- 点推定法 — 標本に基づいてひとつの値で母数の値を推定
- 区間推定法 —- 推定したい母数が「ある確率で入っている」区間を求めること．たとえていえば、多く見積もればいくつ，小さく見積もればいくつといった具合

点推定法から考えていく！

点推定の考えた方

6 個の個体からなる母集団とその特性値を考える：

仮想的な母集団

個体番号	値
A	148
B	160
C	159
D	153
E	151
F	140

母集団分布の母平均と母分散

```
> p1<-c(148,160,159,153,151,140)
> p1
[1] 148 160 159 153 151 140
> mean(p1)              # 母平均
[1] 151.8333
> sum((p1-mean(p1))**2)/length(p1)
[1] 45.80556            # 母分散
> var(p1)
[1] 54.96667
> sum((p1-mean(p1))**2)/(length(p1)-1)
[1] 54.96667
>
```

⋆ この母集団から標本数 $n = 4$ の標本を「無作為非復元抽出法」で抽出する場合には,

$$ {}_6C_4 = {}_6C_2 = \frac{6 \times 5}{2 \times 1} = 15 $$

の標本の取り出し方がある.

	標本	X_1	X_2	X_3	X_4	標本平均の値
1	A, B, C, D	148	160	159	153	155.000
2	A, B, C, E	148	160	159	151	154.500
3	A, B, C, F	148	160	159	140	151.750
4	A, B, D, E	148	160	153	151	153.000
5	A, B, D, F	148	160	153	140	150.250
6	A, B, E, F	148	160	151	140	149.750
7	A, C, D, E	148	159	153	151	152.750
8	A, C, D, F	148	159	153	140	150.000
9	A, C, E, F	148	159	151	140	149.500
10	A, D, E, F	148	153	151	140	148.000
11	B, C, D, E	160	159	153	151	155.750
12	B, C, D, F	160	159	153	140	153.000
13	B, C, E, F	160	159	151	140	152.250
14	B, D, E, F	160	153	151	140	152.500
15	C, D, E, F	159	153	151	140	150.750
	標本平均の平均					151.8333

⋆ 表はすべての標本のパターンとそれぞれの観測値に基づく標本平均の値を列記したものである．

⋆ 実際には，無作為抽出で表の中の一組の値が観測される．

⋆ 母集団の（母）平均 151.8333 の代用品として，観測された標本の値に基づく平均値をもちいたとき，標本に基づいて計算された平均値を母平均の推定値という．

⋆ 注意：母集団分布の特性値（正確には母集団分布を区別するもの）を母数ということにする．したがって，母平均も母集団分布の母数の例である．

⋆ 母集団分布の母数を標本により求められた値により代用することを「点推定」という！

⋆ 一般に，推定対象の母数を θ とかく．観測値を $x_1, x_2, \ldots, x_n$ としたとき，θ の推定値は入力 $x_1, x_2, \ldots, x_n$ に対してもとめることができるので，θ の推定値を

$$\widehat{\theta} := f(x_1, x_2, \ldots, x_n)$$

と書くことにする．

⋆　標本の値は選ばれた標本の値に依存し変化するものなので，標本の値 $x_1, x_2, \ldots, x_n$ は確率変数 $X_1, X_2, \ldots, X_n$ の実現値と解釈する．

⋆ 標本は母集団から無作為に抽出されるので，確率変数 $X_1, X_2, \ldots, X_n$ は

- 互いに独立
- 各 $X_i\,(i = 1, 2, \ldots, n)$ は母集団分布に従う

と考える．

⋆ 関数 $f(x_1, x_2, \ldots, x_n)$ の入力に $X_1, X_2, \ldots, X_n$ を入れたものを θ の推定量という．

$$\widehat{\theta} := f(X_1, X_2, \ldots, X_n)$$

⋆ θ の推定量 $\widehat{\theta}$ と推定値 $\widehat{\theta}$ を区別するために，$\widehat{\theta}(X_1, X_2, \ldots, X_n)$ と $\widehat{\theta}(x_1, x_2, \ldots, x_n)$ とかく．

⋆ θ の推定値 $\widehat{\theta}(x_1, x_2, \ldots, x_n)$ はθ の推定量 $\widehat{\theta}(X_1, X_2, \ldots, X_n)$ の実現値とみなすことができる．

⋆ θ の推定量 $\widehat{\theta}(X_1, X_2, \ldots, X_n)$ は確率変数なので，分布をもつ．これを $\widehat{\theta}(X_1, X_2, \ldots, X_n)$ の**標本分布**という．

⋆ 推定量の精度は標本分布に基づいて議論する．

母集団の個体の値：母集団分布

```
> x<-round(rnorm(100,100,5))
> x
  [1]  104 103  97 105  95 102 101 105 104 105
 [11] 102 104 104 107 105 103  96 100  93 104
 [21]  97 103 102  96  95 107 101 101  98  96
 [31] 110  98  99  98  97  95  97 101 101 101
 [41]  95  96 103 101 101 103  97  97  95 107
 [51] 112 108  93  96 101 103  96 101  95  99
 [61] 100 108  98 112  92  99 105  95  98  92
 [71] 100  95  97  98 108  96  96 106 100  91
 [81]  97  96  99  98 100 101  96 109 101 111
 [91]  104  99  98 102  99  95  97 104  98  87
 > mean(x)
[1] 100.08
> var(x)
[1] 22.90263
> var(x)*(length(x)-1)/length(x)
[1] 22.6736
>
```

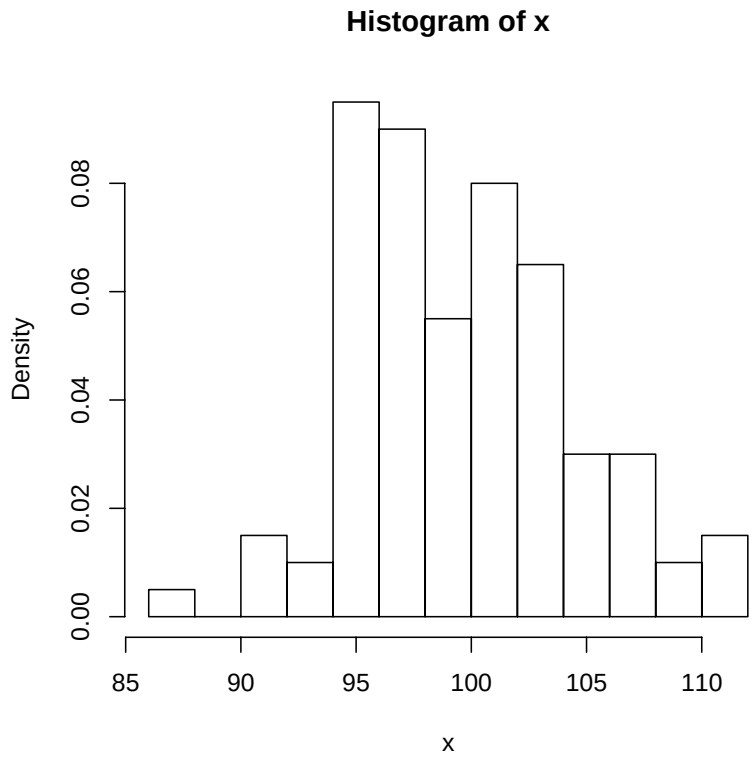

Figure 1: 母集団分布（母平均=100.08, 母分散=22.67）

標本平均と標本中央値の標本分布

```
> no.of.rep<-1000
> sam.mean<-rep(0,no.of.rep)
> sam.median<-rep(0,no.of.rep)
> for (i in 1:no.of.rep){
+ sam.mean[i]<-mean(sample(x,replace=T,10))
+ sam.median[i]<-median(sample(x,replace=T,10))
+ }
> par(mfrow=c(2,1))
> hist(sam.mean,freq=F,nclass=20)
> hist(sam.median,freq=F,nclass=20)
> mean(sam.mean)
[1] 100.0783
> var(sam.mean)
[1] 2.001260
> mean(sam.median)
[1] 99.7225
> var(sam.median)
[1] 3.050294
```

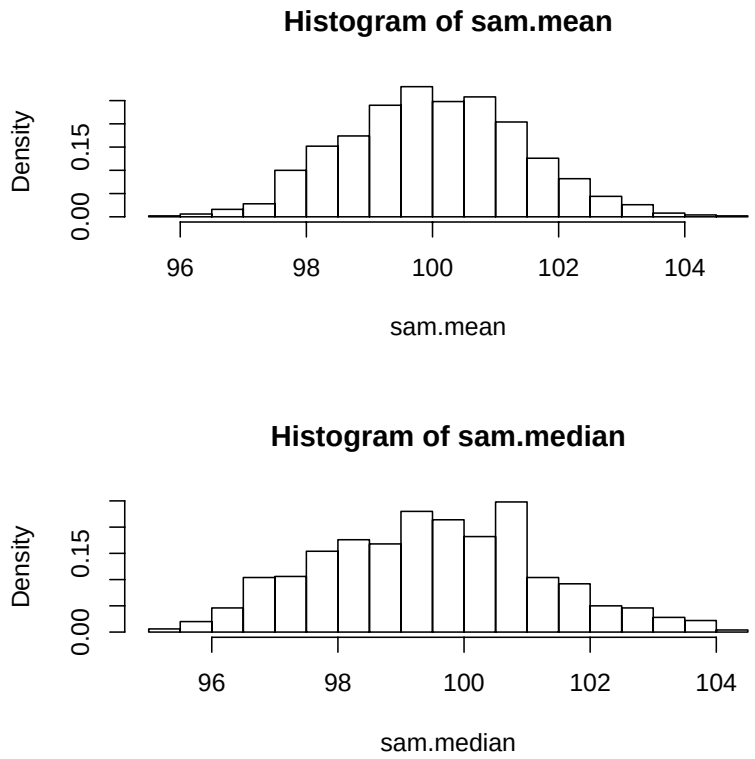

Figure 2: 上段が標本平均の標本分布（平均=100.0783, 分散=2.00）・下段が標本中央値の標本分布（平均=99.7225, 分散=3.05）

ここまでのまとめ

- 母集団・母集団分布・母数（母平均と母分散）・標本という用語を理解しましたか？
- 点推定法をりかいしましたか？
- 推定値と推定量の違いがわかりましたか？
- 推定量の標本分布とはどんな概念か理解しましたか？

推定量の望ましい性質

- 不偏性 — 推定量が偏った形で母数を推定しなこと！
- 一致性 — 標本の数を無限大にしたとき，推定量が真の母数の値に近づくこと.
- 有効性：最良線形不偏推定量 — 線形不偏推定量中で標本分布の分散が最小のもの.

不偏性について

標本の実現値

$$X_1 = x_1,\, X_2 = x_2,\, \ldots,\, X_n = x_n$$

に基づいた推定値$\widehat{\theta}(x_1, x_2, \ldots, x_n)$ は母数の値 θ より大きめの値になったり，小さめな値となる可能性がある．

⋆ 仮想的な母集団

個体番号	値
A	148
B	160
C	159
D	153
E	151
F	140

を思い出そう．推定したい母数を母平均 = 151.8333 とする．

	標本	x_1	x_2	x_3	x_4	標本平均の値
1	A, B, C, D	148	160	159	153	155.000
2	A, B, C, E	148	160	159	151	154.500
3	A, B, C, F	148	160	159	140	151.750
4	A, B, D, E	148	160	153	151	153.000
5	A, B, D, F	148	160	153	140	150.250
6	A, B, E, F	148	160	151	140	149.750
7	A, C, D, E	148	159	153	151	152.750
8	A, C, D, F	148	169	153	140	150.000
9	A, C, E, F	148	169	151	140	149.500
10	A, D, E, F	148	153	151	140	148.000
11	B, C, D, E	160	159	153	151	155.750
12	B, C, D, F	160	159	153	140	153.000
13	B, C, E, F	160	159	151	140	152.250
14	B, D, E, F	160	153	151	140	152.500
15	C, D, E, F	159	153	151	140	150.750
	標本平均の平均					151.8333

つぎの等式が成立している：

母平均 = 標本平均の標本分布の平均

すわわち，推定値が母平均に比べて小さくなったり大きくなったりするが，平均してみると母平均に一致している！

不偏推定量の定義

母数 θ の推定量 $\widehat{\theta}$ が θ の不偏推定量であるとは，

$$\mathbb{E}[\widehat{\theta}] = \theta$$

が成立しているときをいう.

この式の意味は,「母数 θ の推定量 $\widehat{\theta}$ の標本分布の平均は母数 θ に一致する」となる！

$\star$ 一般に（母平均が存在すれば），標本平均は母平均の不偏推定量である！

なぜならば，$\mathbb{E}[X_i] = \theta\,(i = 1, 2, \ldots, n)$ なので，

$$\begin{aligned} \mathbb{E}[\bar{X}_n] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] = \theta \end{aligned}$$

母分散の不偏推定量：母平均が既知のとき

$X_1, X_2, \ldots, X_n\,(n \geq 2)$ は独立同一分布に従い，$\mathbb{E}[X_1] = \theta$, $\mathbb{VAR}[X_1] = \sigma^2$ とする．ここで，θ の値はわかっているものとする．

このとき，σ^2 の推定量として

$$\widehat{\sigma}^2 = c\sum_{i=1}^{n}(X_i - \theta)^2$$

を考えよう．ただし，c は正の定数とする．

このとき,

$$\begin{aligned}
\mathbb{E}[\hat{\sigma}^2] &= \mathbb{E}[c\sum_{i=1}^{n}(X_i-\theta)^2] \\
&= c\sum_{i=1}^{n}\mathbb{E}[(X_i-\theta)^2] \\
&= c\sum_{i=1}^{n}\mathbb{E}[(X_i-\mathbb{E}[X_i])^2] \\
&= c\sum_{i=1}^{n}\mathbb{VAR}[X_i] = nc\sigma^2
\end{aligned}$$

よって,

$$\mathbb{E}[\hat{\sigma}^2] = \sigma^2 \Longleftrightarrow c = \frac{1}{n}$$

母平均が既知のときの母分散の不偏推定量

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \theta)^2$$

が成立しているときをいう.

母分散の不偏推定量：母平均が未知のとき

θ の値が未知のとき，θ の推定量 $\bar{X}_n$ で代用すると母分散の推定量

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

これは母分散の不偏推定量だろうか？そのためには期待値

$$\mathbb{E}[\widehat{\sigma}^2]$$

を評価すればよい.

$$
\begin{aligned}
\mathbb{E}[\widehat{\sigma}^2] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X}_n)^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\{(X_i-\theta)+(\theta-\bar{X}_n)\}^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\left\{\sum_{i=1}^{n}(X_i-\theta)^2+\sum_{i=1}^{n}(\theta-\bar{X}_n)^2+2(\theta-\bar{X}_n)\sum_{i=1}^{n}(X_i-\theta)\right\}\right] \\
&= \mathbb{E}\left[\frac{1}{n}\left\{\sum_{i=1}^{n}(X_i-\theta)^2+n(\theta-\bar{X}_n)^2+2n(\theta-\bar{X}_n)(\bar{X}_n-\theta)\right\}\right] \\
&= \mathbb{E}\left[\frac{1}{n}\left\{\sum_{i=1}^{n}(X_i-\theta)^2-n(\theta-\bar{X}_n)^2\right\}\right]
\end{aligned}
$$

よって,

$$\begin{aligned}
\mathbb{E}[\widehat{\sigma}^2] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\theta)^2-(\theta-\bar{X}_n)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(X_i-\theta)^2]-\mathbb{E}[(\theta-\bar{X}_n)^2] \\
&= \frac{1}{n}\times n\mathbb{VAR}[X_i]-\mathbb{VAR}[\bar{X}_n] \\
&= \sigma^2-\frac{\sigma^2}{n}=\frac{n-1}{n}\sigma^2
\end{aligned}$$

なぜならば,

$$\mathbb{E}[\bar{X}_n] = \theta, \qquad \mathbb{VAR}[\bar{X}_n]=\frac{\sigma^2}{n}$$

であった.

母平均が既知のときの母分散の不偏推定量

$$\widehat{\sigma}_U^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2, \qquad \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

としたとき,

$$\mathbb{E}[\widehat{\sigma}_U^2] = \sigma^2$$

が成立しているときをいう.

ここまでのまとめ

- 推定量の不偏性という概念を理解しましたか？
  - 標本平均は母平均の不偏推定量である.
  - 標本分散の推定量を不偏にするために $\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ を $n-1$ でわっている.

データの値を $x_1, x_2, \ldots, x_n$ とし，それを小さいものから並べなおしたものを

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

と書くことにする.

たとえば，データの値が $x_1 = 4, x_2 = 1, x_3 = 3, x_4 = 5, x_5 = -1$ であれば，

$$x_{(1)} = -1, x_{(2)} = 1, x_{(3)} = 3, x_{(4)} = 4, x_{(5)} = 5$$

標本中央値を

$$\mathsf{median} = \begin{cases} x_{(n+1)/2} & n \text{ が奇数} \\ \dfrac{(x_{(n/2)} + x_{(n/2+1}}{2} & n \text{ が偶数} \end{cases}$$

$x_1 = 4, x_2 = 1, x_3 = 3, x_4 = 5$ のとき，

$$x_{(1)} = 1, x_{(2)} = 3, x_{(3)} = 4, x_{(4)} = 5$$

となるので，標本中央値は

$$\frac{3+4}{2} = 3.5$$

問題 1

母集団

```
> x<-round(rt(10000,5))*3+20
```

で生成する．

(1) この母集団分布の母平均と母分散およびヒストグラムを作成せよ．（mejiro-hanako-071221-median1.pdf）

(2) 標本の大きさを誕生月と 7 のおおきい方として，標本中央値を 5000 個生成し，そのヒストグラム（標本中央値の標本分布）とその平均値を求め，標本中央値が母平均の不偏推定量かどうかを推論（近い値であれば、不偏性をもつと予想してよい）せよ．ヒストグラムを mejiro-hanako-071221-median2.pdf とせよ．

ヒント

```
> no.of.rep<-5000
> sam.median<-rep(0,no.of.rep)
> for (i in 1:no.of.rep){
+ sam.median[i]<-median(sample(x,replace=T,??))
+ }
> hist(sam.median,freq=F,nclass=20)
> mean(sam.median)
```

- それぞれの問題の解説を mejiro-hanako-071221.txt に書け．
- 締め切りは 2007 年 12 月 21 日（金）13 時

問題 2

母集団

```
> z<-rt(2000,10)*5-5
> y<-rt(3000,5)*3+20
> x<-c(y,z)          母集団の個体の値がオブジェクト
```

で生成する.

(1) この母集団分布の母平均と母分散およびヒストグラムを作成せよ．母集団分布がどのような形をしているか？（mejiro-hanako-071221-median3.pdf）

(2) 標本の大きさを誕生月と 7 のおおきい方として，標本平均値を 5000 個生成し，そのヒストグラム（標本平均値の標本分布）とその平均値を求め，標本平均値が母平均の不偏推定量かどうかを推論（近い値であれば、不偏性をもつと予想してよい）せよ．ヒストグラムを mejiro-hanako-071221-median4.pdf とせよ.

(2) のヒント

```
> no.of.rep<-5000
> sam.mean<-rep(0,no.of.rep)
> for (i in 1:no.of.rep){
+ sam.mean[i]<-mean(sample(x,replace=T,??))
+ }
> hist(sam.mean,freq=F,nclass=20)
> mean(sam.mean)
```

(3) 標本の大きさを誕生月と 7 のおおきい方として，標本中央値を 5000 個生成し，そのヒストグラム（標本中央値の標本分布）とその平均値を求め，標本中央値が母平均の不偏推定量かどうかを推論（近い値であれば、不偏性をもつと予想してよい）せよ．ヒストグラムを mejiro-hanako-071221-median5.pdf とせよ．

(3) のヒント

```
> no.of.rep<-5000
> sam.median<-rep(0,no.of.rep)
> for (i in 1:no.of.rep){
+ sam.median[i]<-median(sample(x,replace=T,??))
+ }
> hist(sam.median,freq=F,nclass=20)
> mean(sam.median)
```

- それぞれの問題の解説を mejiro-hanako-071221.txt に書け.
- 締め切りは 2007 年 12 月 21 日（金）13 時

問題 3

母集団

```
> x<-rf(5000,8,15)*{(誕生月/2+2)の平方根}
```

で生成する．

(1) この母集団分布の母分散およびヒストグラムを作成せよ．母集団分布がどのような形をしているかを述べよ．

ヒント

```
> hist(x,freq=F,nclass=??)
> var(x)*(length(x)-1)/length(x)
```

(2) $\sum_{i=1}^{n}(X_i - \bar{X}_n)^2/(n-1)$ の値を 5000 個生成し，そのヒストグラム（不偏分散の標本分布）と平均値を求め，$\sum_{i=1}^{n}(X_i - \bar{X}_n)^2/(n-1)$ の母分散の不偏推定量かどうかを（近い値であれば、不偏性をもつと予想してよい）せよ．ただ

し，標本の大きさを誕生月と 5 のおおきい方とする．ヒストグラムを mejiro-hanako-071221-median6.pdf とせよ．

(2) のヒント

```
> no.of.sample<-5000
> sam.var<-rep(0,no.of.sample)
> for (i in 1:no.of.sample){
+ z<-sample(x,replace=T,??)
+ sam.var[i]<-sum((z-mean(z))**2/(length(z)-1))
+ }
> hist(sam.var,freq=F,nclass=20)
> mean(sam.var)
```

(3) $\sum_{i=1}^{n}(X_i - \bar{X}_n)^2/n$ の値を 5000 個生成し，そのヒストグラム（不偏分散の標本分布）と平均値を求め，$\sum_{i=1}^{n}(X_i - \bar{X}_n)^2/n$ の母分散の不偏推定量かどうかを（近い値であれば、不偏性をもつと予想してよい）せよ．ただし，標本の大きさを誕生月と 5 のおおきい方とする．ヒストグラムを mejiro-hanako-071221-median7.pdf とせよ．

(3) ヒント

```
> no.of.sample<-5000
> sam.var2<-rep(0,no.of.sample)
> for (i in 1:no.of.sample){
+ z<-sample(x,replace=T,??)
+ sam.var2[i]<-sum((z-mean(z))**2/length(z)
+ }
> hist(sam.var2,freq=F,nclass=20)
> mean(sam.var2)
```

- それぞれの問題の解説を mejiro-hanako-071221.txt に書け.
- 締め切りは 2007 年 12 月 21 日（金）13 時

一致性

標本数 n を無限大にもっていったときに，$\widehat{\theta}$ は θ に確率的に近づいてほしい！

一致性の定義

すわわち，任意の正数 $\epsilon > 0$ に対し，

$$\lim_{n\to\infty} \mathbb{P}(|\widehat{\theta} - \theta| > \epsilon) = 0$$

標本数をおおきくしたときの標本分布の変化：母集団の作成

```
> x<-round(rnorm(10000,100,5))
> mean(x)
[1] 100.0167
> var(x)*(length(x)-1)/length(x)
[1] 24.91502
```

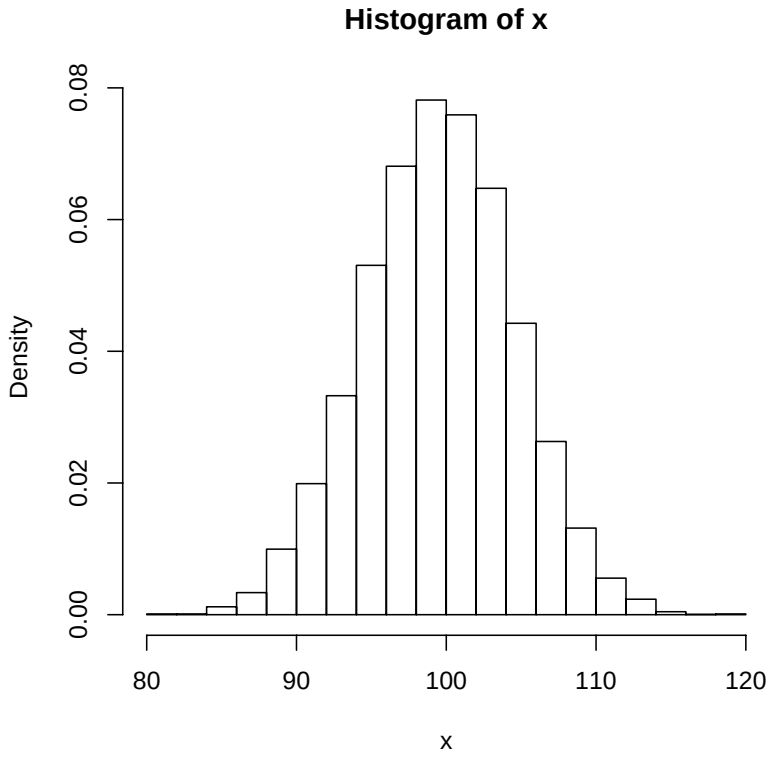

Figure 3: 母平均=100.0167, 分散=24.91502

標本数をおおきくしたときの標本分布の変化

```
> no.of.rep<-1000
> sam.mean10<-rep(0,no.of.rep)
> sam.mean20<-rep(0,no.of.rep)
> sam.mean50<-rep(0,no.of.rep)
> sam.mean100<-rep(0,no.of.rep)
> for (i in 1:no.of.rep){
+ sam.mean10[i]<-mean(sample(x,replace=T,10))
+ sam.mean20[i]<-mean(sample(x,replace=T,20))
+ sam.mean50[i]<-mean(sample(x,replace=T,50))
+ sam.mean100[i]<-mean(sample(x,replace=T,100))
+ }
> hist(sam.mean10,freq=F,nclass=20,main=paste("n=10"))
> hist(sam.mean20,freq=F,nclass=20,main=paste("n=20"))
> hist(sam.mean50,freq=F,nclass=20,main=paste("n=50"))
> hist(sam.mean100,freq=F,nclass=20,main=paste("n=100"))
```

標本数をおおきくしたときの標本分布の変化

```
> no.of.rep<-1000
> sam.mean1000<-rep(0,no.of.rep)
> sam.mean10000<-rep(0,no.of.rep)
> sam.mean100000<-rep(0,no.of.rep)
> sam.mean1000000<-rep(0,no.of.rep)
> for (i in 1:no.of.rep){
+ sam.mean1000[i]<-mean(sample(x,replace=T,1000))
+ sam.mean10000[i]<-mean(sample(x,replace=T,10000))
+ sam.mean100000[i]<-mean(sample(x,replace=T,100000))
+ sam.mean1000000[i]<-mean(sample(x,replace=T,1000000))
+ }
> hist(sam.mean1000,freq=F,nclass=20,main=paste("n=1000"))
> hist(sam.mean10000,freq=F,nclass=20,main=paste("n=10000"))
> hist(sam.mean100000,freq=F,nclass=20,main=paste("n=100000"))
> hist(sam.mean1000000,freq=F,nclass=20,main=paste("n=1000000"))
> hist(sam.mean1000000,freq=F,nclass=20,main=paste("n=1000000"))
> par(mfrow=c(1,1))
```

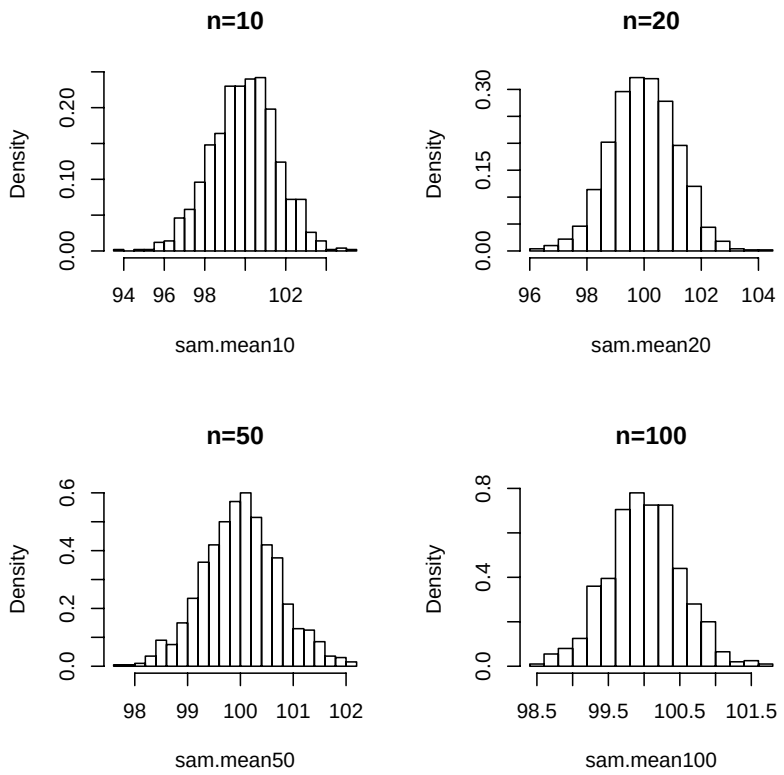

Figure 4: $n = 10, 20, 50, 100$ の標本平均の標本分布

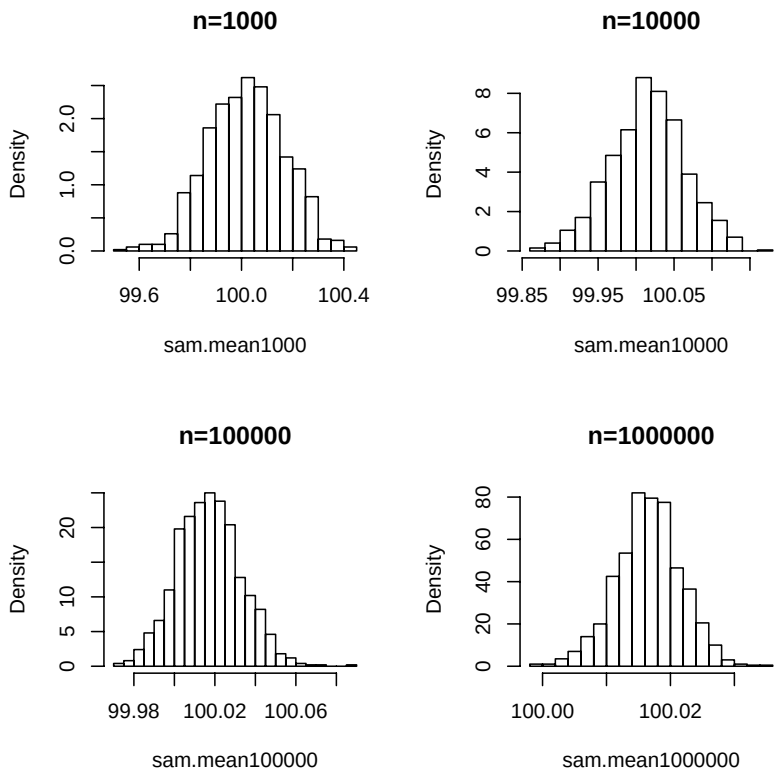

Figure 5: $n = 1000, 10000, 100000, 1000000$ の標本平均の標本分布

有効性：最良線形不偏推定量

母平均を標本数 $n = 4$ の標本 X_1, X_2, X_3, X_4 で推定することを考える．推定量として，

$$\begin{aligned}
\widehat{\theta}_1 &= \frac{X_1}{1} \\
\widehat{\theta}_2 &= \frac{X_1 + X_2}{2} \\
\widehat{\theta}_3 &= \frac{X_1 + X_2 + X_3}{3} \\
\widehat{\theta}_4 &= \frac{X_1 + X_2 + X_3 + X_4}{4}
\end{aligned}$$

これらはすべて θ の不偏推定量：

$$\mathbb{E}[\widehat{\theta}_i] = \theta \qquad (i = 1, 2, 3, 4)$$

しかし，$\mathbb{VAR}[X_1] = \sigma^2$ とすれば，

$$\begin{aligned} \mathbb{VAR}[\widehat{\theta}_1] &= \frac{\sigma^2}{1}, \qquad \mathbb{VAR}[\widehat{\theta}_2] = \frac{\sigma^2}{2} \\ \mathbb{VAR}[\widehat{\theta}_3] &= \frac{\sigma^2}{3} \\ \mathbb{VAR}[\widehat{\theta}_4] &= \frac{\sigma^2}{4} \end{aligned}$$

よって，$\widehat{\theta}_4$ の分散が一番小さい！

一般に，$X_1, X_2, \ldots, X_n$ に基づいて母平均 θ を推定することを考える．ここで，$\mathbb{E}[X_i] = \theta$, $\mathbb{VAR}[X_i] = \sigma^2\,(i = 1, 2, \ldots, n)$ とする．$c_1, c_2, \ldots, c_n$ を定数とする．

線形推定量

$$\widehat{\theta}_{\mathbf{c}} = c_1X_1 + c_2X_2 + \cdots + c_nX_n, \qquad \mathbf{c} = (c_1, c_2, \ldots, c_n)$$

$c_1 + c_2 + \cdots + c_n = 1$ のとき，線形推定量はθ の不偏推定量となる：

$$\mathbb{E}[\widehat{\theta}_{\mathbf{c}}] = \theta$$

なぜならば，

$$\begin{aligned}
\mathbb{E}[\widehat{\theta}_{\mathbf{c}}] &= \mathbb{E}[c_1X_1 + c_2X_2 + \cdots + c_nX_n] \\
&= c_1\mathbb{E}[X_1] + c_2\mathbb{E}[X_2] + \cdots + c_n\mathbb{E}[X_n] \\
&= c_1\theta + c_2\theta + \cdots + c_n\theta \\
&= \theta(c_1 + c_2 + \cdots + c_n) = \theta
\end{aligned}$$

一方，$\widehat{\theta}_{\mathbf{c}}$ の分散は

$$\begin{aligned}
\mathbb{VAR}[\widehat{\theta}_{\mathbf{c}}] &= \mathbb{VAR}[c_1X_1 + c_2X_2 + \cdots + c_nX_n] \\
&= \mathbb{VAR}[c_1X_1] + \mathbb{VAR}[c_2X_2] + \cdots + \mathbb{VAR}[c_2X_n] \qquad \text{(独立性より)} \\
&= c_1^2\mathbb{VAR}[X_1] + c_2^2\mathbb{VAR}[X_2] + \cdots + c_n^2\mathbb{VAR}[c_2X_n] \\
&= (c_1^2 + c_2^2 + \cdots + c_n^2)\sigma_2^2
\end{aligned}$$

しかし，

$$\begin{aligned} c_1^2 + c_2^2 + \cdots + c_n^2 &= \left(c_1 - \frac{1}{n}\right)^2 + \left(c_2 - \frac{1}{n}\right)^2 + \cdots + \left(c_n - \frac{1}{n}\right)^2 \\ &\quad + \frac{2}{n}(c_1 + c_2 + \cdots + c_n) - \frac{1}{n} \\ &= \left(c_1 - \frac{1}{n}\right)^2 + \left(c_2 - \frac{1}{n}\right)^2 + \cdots + \left(c_n - \frac{1}{n}\right)^2 + \frac{1}{n} \end{aligned}$$

よって，

$$\mathbb{VAR}[\widehat{\theta}_{\mathbf{c}}] \geq \frac{\sigma^2}{n}$$

等号成立は

$$c_1 = \frac{1}{n},\ c_2 = \frac{1}{n},\ \ldots,\ c_n = \frac{1}{n}$$

最良線形推定量

$$\widehat{\theta}_{\mathbf{c}} = c_1 X_1 + c_2 X_2 + \cdots + c_n X_n, \qquad \mathbf{c} = (c_1, c_2, \ldots, c_n)$$

は $c_1 + c_2 + \cdots + c_n = 1$ のとき，θ の不偏推定量.
さらに,

$$\mathbb{VAR}[\widehat{\theta}_{\mathbf{c}}] \geq \mathbb{VAR}[\bar{X}_n]$$

すなわち,

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

は最良線形不偏推定量！

ここまでのまとめ

- 母集団と標本分布を理解しましか？
- 母平均と標本平均の違いがわかりますか？
- 推定量と推定値の違いがわかりますか？
- 点推定量：性質と推定量の導出について
  - 不偏推定量の概念を理解しましたか？
  - 有効推定量：最良線形不偏推定量の概念を理解しましたか？