データ解析基礎論a1 講義02

**資料の見方**：

　重要度１：データ解析基礎論a1~a2においては「気にしなくても良い」レベル

　重要度２：知っておいて欲しいレベル

重要度３：試験で問われる可能性があるレベル

1. Rの初歩的な処理２　[重要度3]
   1. 条件＋α（重要度３）

データなどから特定の条件を満たすものを取り出す。

#まずデータを取り入れます：

dat<-read.csv("http://www.matsuka.info/data\_folder/datWA01.txt",

header=T);

# head(data\_set)で、データが正しく取り入れらたか確認します。

> head(dat)

shoesize h gender

1 27.0 181.4 M

2 26.5 170.8 M

3 27.5 182.3 M

4 26.5 166.8 M

5 23.5 153.2 F

6 23.0 151.6 F

データには足のサイズ（shoesize）身長（h）と性別(gender)があります。ここで男性（女性）のデータを取り出したい場合、講義の論理演算を適用すると：

dat\_male = dat[which(dat$gende=="M"), ]

> head(dat\_male)

shoesize h gender

1 27.0 181.4 M

2 26.5 170.8 M

3 27.5 182.3 M

4 26.5 166.8 M

13 26.0 173.0 M

14 26.5 180.0 M

のように、男性だけのデータを取り出すことが可能です。

# 女性のデータ

dat$shoesize[dat$gender == "F"]

# 身長180以上のデータ

dat\_ge180 = dat[dat$h >= 180, ]

1.1.1 which

使用方法：

which(expression…)

Rの実行例

> a=1:10

[1] 1 2 3 4 5 6 7 8 9 10

> which(a<5)

[1] 1 2 3 4

#aの1,2,3,4番目の要素が５より小さい

> b=10:1

[1] 10 9 8 7 6 5 4 3 2 1

> which(b<5)

[1] 7 8 9 10

#bの7,8,9,10番目の要素が５より小さい

#講義１の実習２から4月のみを抽出

dat <- data.frame(subjID = rep(1:5,2),

Month = c(rep("april",5),rep("may",5)),

Succeeded = rep(c("Yes","No"),5))

dat\_april = dat[dat$Month=="april",]

> head(dat\_april)

subjID Month Succeeded

1 1 april Yes

2 2 april No

3 3 april Yes

4 4 april No

5 5 april Yes

1. 記述統計　[定義は重要度3、数式は重要度２]

2.1 母数と統計量（重要度３）

* 母集団(population)
  + 定義：対象全体
  + 母数(parameter) – 母集団の変数の分布の性質を表す値（一般的にギリシャ文字で表記される）
* 標本(sample)
  + 定義：母集団から抽出された一部の対象（サンプル）
  + 統計量(statistics) – サンプルデータから推定する母数の値（一般的にローマ字で表記される）

2.2 平均値 - mean（重要度３）

期待値：予想される数値。特定の変数の（他の情報がない場合の）期待値は、その変数の平均値となる。

母集団・母分布の平均の定義

・離散的変数の場合

・連続的変数の場合

*xi*: 確率変数の実数

*pi*: *xi* が起こる確率

*f(x)*:確率密度

母分布を知ることはまず無く、一般的には計測されたデータ（サンプル）から母分布の平均を以下の式から推定する：

例:X={1,2,3,4,5}, mean(X) = 3

Rの実行例　– mean(x)

使用方法：

mean(var)

x = 1:5

> mean(x)

[1] 3

2.2.1 平均値に関する役立つ定理（重要度２）

X, Y: 変数

a, b, c:　定数

E( ): 期待値（平均値）

x

* E(cX)=cE(X)

例：身長をcmからmに変更した場合

* E(X+c)=E(X)+c

例：全員が厚み3cmの下駄を履いた時の身長

* E(X + Y) = E(X) + E(Y)

例：英語と数学の合計点の平均

2.3 分散と標準偏差 – variance & standard deviation（重要度３）

分散: 平均を中心とした変数のばらつきの２乗の期待値

集団・母分布の分散の定義

または：

とも表せる。

分散は、ばらつきの２乗となっていて、ばらつきが大きく感じることが多く、ルートをとってもともとのスケールに戻したものが標準偏差である。

2.3.1 分散に関する役立つ定理（重要度２）

X, Y: 変数;

a, b, c:定数

* var(cX)=c2var(X)

例：身長をcmからmにした場合

* var(X+c)=var(X)

例：全員が厚み3cmの下駄を履いた時の身長

* var(aX + b)=a2var(X)

例：得点を偏差値に変更した場合

* var(X + Y) = var(X) + var(Y)+2\*cov(X,Y)

例：英語と数学の合計点の分散

ここでcov(X,Y)はXとYの共分散を示す。

2.3.2分散の推定値（重要度３）

平均値と同様に、母分布を知ることはまず無く、一般的には計測されたデータ（サンプル）から母分布の分散を以下の式から推定する（最小二乗法）：

平均値の統計量はnで割っていたが、分散の統計量はn-1で割るのは以下の説明・式で示すように不偏推定値であるからです。

2.3.3 分散の不偏推定値の説明（重要度１―ほぼ０）

以上のことからN-1を用いることによって、分散の期待値は母分散と一致することが示された。以下、計算機シミュレーションの結果を用いて体験する。

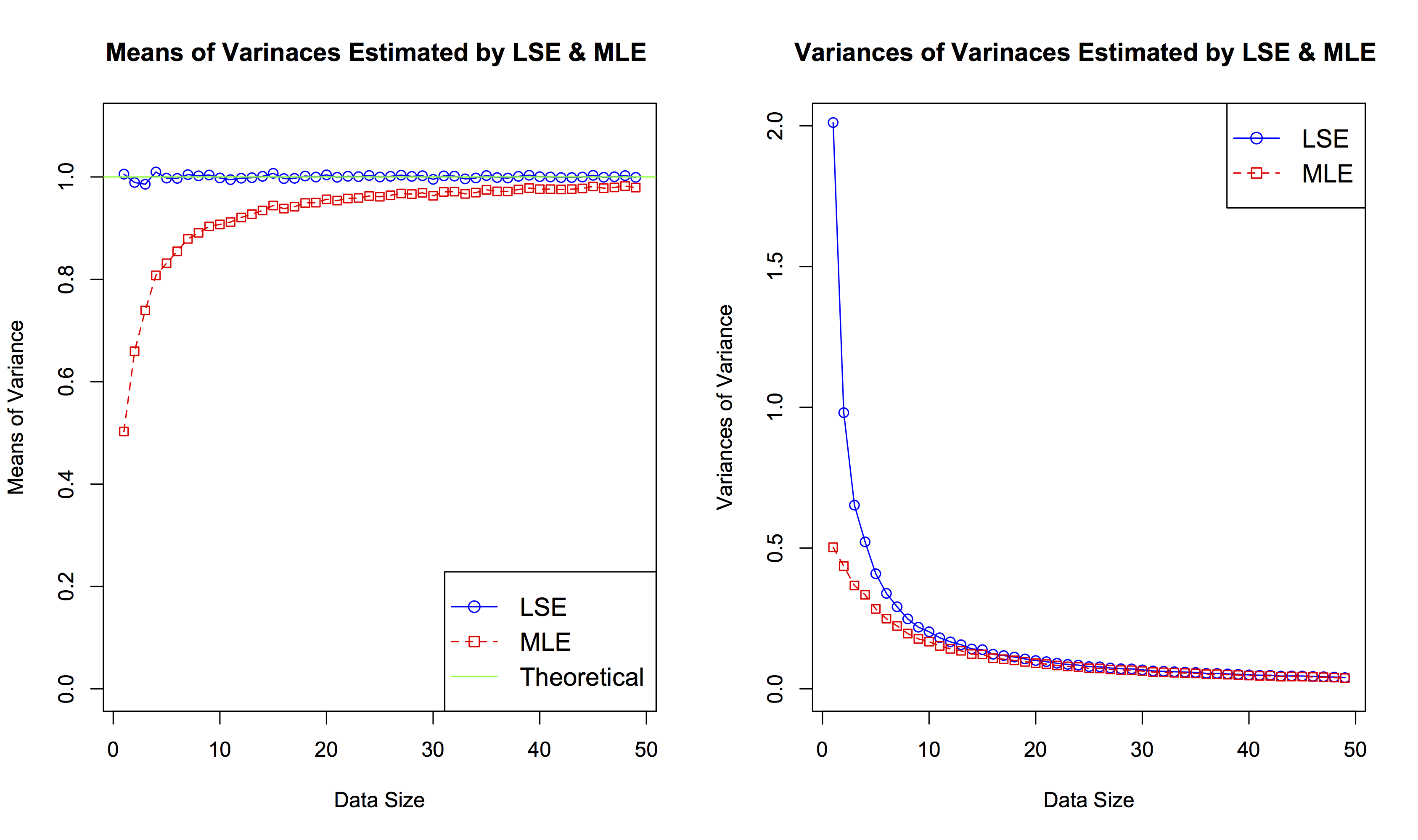


図1．左：x軸はデータ数、Y軸は推定された分散の平均。LSEは最小二乗法（n-1で割る、不偏分散）、MLEは最尤法(nで割る)による推定。最小二乗法の平均値は理論値とほぼ一致するが、最尤法ではデータ数が少ない場合は過小に推定されている。右：x軸はデータ数、Y軸は推定された分散の分散。最小二乗法ではデータ数が少ないほど推定値が安定していない（分散が多い）が、最尤法ではデータ数が少無くとも推定値は安定している。



図２.LSEとMLEによる分散の推定値の分布

2.3.3 var(x) & sd(x)

Rの実行例　– var(x), sd(x)

使用方法：

var(variable), sd(variable)

x = 1:10

> var(x)

[1] 9.166667

> sd(x)

[1] 3.02765

# 定理の検証

* var(cX)=c2var(X)

例：身長をcmからmにした場合

> x2 = x\*10

> var(x2)

[1] 916.6667

> var(x)\*10^2

[1] 916.6667

* var(X+c)=var(X)

例：全員が厚み3cmの下駄を履いた時の身長

> x3 = x +5

> var(x3)

[1] 9.166667

> var(x)

[1] 9.166667

* var(aX + b)=a2var(X)

> x4 = x\*3 + 10

> var(x4)

[1] 82.5

> var(x)\*3^2

[1] 82.5

2.4 共分散と相関係数 – covariance & correlation（重要度３）

２つの変数が同時にどれだけ変化するか

データポイントAにおいて、変数Xがその平均値から離れた場合、変数YがどれだけYの平均値から離れるか。（また、その逆）例：足のサイズと身長の共分散の場合　―　足のサイズの平均値（２５.５cm）でありある人物の足のサイズが２７.５cmであった場合、その人物の身長は身長の平均値（１６５cm）からどの程度異なるかに対する期待値。

母集団の共分散の定義：

同一の変数の共分散は分散です。

共分散も、ばらつきの２乗となり、また2つの変数でスケールの違い（足のサイズはcm、身長はm）により、直感的に解釈が困難ば場合があるため、最大値が１、最小値が−１となる相関係数に変換することも多い。

相関係数の定義：

上記の相関係数は線形の関係の指標であり、「相関なし　→　独立」は必ずしも正しくない

（ただし「独立　→　相関なし」は正しい）。

2.4.1 共分散に関する役立つ定理（重要度２、式の詳細は重要度１）

* cov(X,Y) = cov(Y,X)
* cov(X,X) = var(X)

* cov(aX, bY) = ab\*cov(X,Y)

* cov(a+X, b+Y) = cov(X,Y)

* cov(X, Y+Z) = cov(X,Y)+cov(X,Z)

cov(ΣX, ΣY)=ΣΣcov(X,Y)

2.4.2 cov(x,y) & cor(x,y)

Rの実行例　– cov(x,y), cor(x)

使用方法：

x = 1:10

y = 10:1

> cov(x,y)

[1] -9.166667

> cor(x,y)

[1] -1