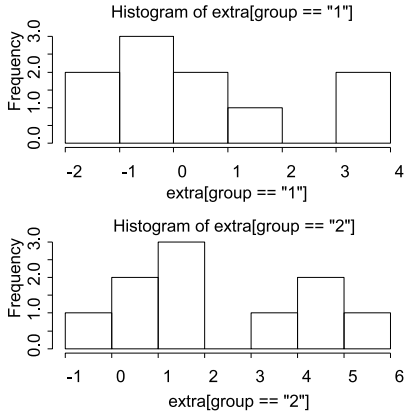
医学情報処理演習第４回「標本統計量と母数推定」 2006 年 10 月 30 日 中澤　港

前回の課題の回答例

sleep データの分布の位置とばらつきの情報を与える記述統計量を計算する前に，図示により分布の様子を確認する。図示の方法はいろいろあるが，とりあえずヒストグラムを作ってみる。

```
attach(sleep)
layout(c(1,2))
hist(extra[group=="1"])
hist(extra[group=="2"])
detach(sleep)
```

により下図ができ，正規分布とは到底言えなそうだとわかる。そこで位置とばらつきの情報としては，平均±不偏標準偏差よりもむしろ中央値±四分位偏差を示すべきと判断される。




```
SIQR <- function(x) { (fivenum(x)[4]-fivenum(x)[2])/2 }
attach(sleep)
tapply(extra,group,median)
tapply(extra,group,SIQR)
detach(sleep)
```

により（tapply は，１番目の引数について，２番目の引数ごとに層別して，３番目の引数で与える関数を適用するという関数なので，group 別の統計量を計算する時に便利），催眠薬１を投与したときの睡眠時間変化の中央値±四分位偏差は 0.35 ± 1.1，催眠薬２投与時の睡眠時間変化の中央値±四分位偏差は 1.75 ± 1.8 とわかる。

標本統計量と母数

前回も軽く説明したように，我々が統計解析で相手にするのは標本データの場合が多く，得られる統計量も標本統計量であるが，本当に知りたいのは母集団の統計量（母数）であるため，標本統計量から母数を推定しなければならない。その仕組みを考えるため，まず標本抽出をシミュレートしてみる。

標本抽出

表 1. 成人男子の身長からなる母集団と母平均 μ，母分散 σ^2 の計算

身長 (cm) (x)	度数	相対度数 $(p(x))$	$x \cdot p(x)$	$(x-\mu)^2 \cdot p(x)$
144	2	0.000002	0.000288	1.151979e−03
145	3	0.000003	0.000435	1.586970e−03
146	6	0.000006	0.000876	2.903942e−03
147	17	0.000017	0.002499	7.496844e−03
148	19	0.000019	0.002812	7.599834e−03
149	56	0.000056	0.008344	2.021553e−02
150	125	0.000125	0.018750	4.049901e−02
151	219	0.000219	0.033069	6.328937e−02
152	463	0.000463	0.070376	1.185248e−01
153	915	0.000915	0.139995	2.058690e−01
154	1609	0.001609	0.247786	3.153541e−01
155	2649	0.002649	0.410595	4.476659e−01
156	4550	0.004550	0.709800	6.551761e−01
157	7214	0.007214	1.132598	8.728592e−01
158	11005	0.011005	1.738790	1.100452e+00
159	16081	0.016081	2.556879	1.302498e+00
160	22098	0.022098	3.535680	1.414195e+00
161	29903	0.029903	4.814383	1.465155e+00
162	39048	0.039048	6.325776	1.405625e+00
163	48312	0.048312	7.874856	1.207694e+00
164	57703	0.057703	9.463292	9.231469e−01
165	66639	0.066639	10.995435	5.996634e−01
166	73332	0.073332	12.173112	2.932638e−01
167	78051	0.078051	13.034517	7.801682e−02
168	79829	0.079829	13.411272	3.828679e−09
169	77866	0.077866	13.159354	7.790011e−02
170	73767	0.073767	12.540390	2.951326e−01
171	66321	0.066321	11.340891	5.969761e−01
172	57993	0.057993	9.974796	9.279896e−01
173	48410	0.048410	8.374930	1.210356e+00
174	39081	0.039081	6.800094	1.407019e+00
175	29967	0.029967	5.244225	1.468475e+00
176	22055	0.022055	3.881680	1.411597e+00
177	15810	0.015810	2.798370	1.280672e+00
178	10875	0.010875	1.935750	1.087548e+00
179	7309	0.007309	1.308311	8.844242e−01
180	4596	0.004596	0.827280	6.618482e−01
181	2726	0.002726	0.493406	4.607095e−01
182	1519	0.001519	0.276458	2.977333e−01
183	939	0.000939	0.171837	2.112812e−01
184	462	0.000462	0.085008	1.182752e−01
185	224	0.000224	0.041440	6.473767e−02
186	128	0.000128	0.023808	4.147301e−02
187	50	0.000050	0.009350	1.805042e−02
188	31	0.000031	0.005828	1.240027e−02
189	14	0.000014	0.002646	6.174129e−03
190	5	0.000005	0.000950	2.420048e−03
191	4	0.000004	0.000764	2.116040e−03
合計	1000000	1.00	$\mu = 167.9998$	$\sigma^2 = 25.09521$
				$\sigma = 5.009512$

100 万人の成人男性の身長からなる母集団が，表 1 のようになっていたとしよう[*1]。

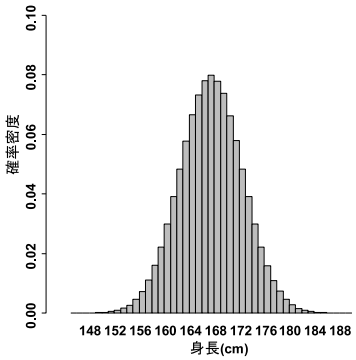


表 1 のように計算すれば，X の平均 $\mu = \sum x \cdot p(x) = 167.9998$，X の分散 $\sigma^2 = 25.09521$ である。ここから 10 個の標本を抽出することを考える。なるべく母集団全体の情報を偏りなく代表させるためには，身長が書かれた 100 万枚のカードを袋に入れてよくかきまぜ，10 枚取り出すことが考えられる。無作為に抽出される最初のカードが x である確率は，$p(x)$ である。この 1 回目の抽出を X_1 であらわす。2 枚目のカードについても同じである（厳密に言えば，1 枚とっては元に戻して再びよくかきまぜて抽出する「復元抽出」でないと，1 枚目に何が出たかに依存して相対度数が変化するため，条件付き確率はまったく同じにはならないが，100 万くらいの大きな母集団からサイズ 10 の標本を取り出すのであれば，「非復元抽出」（1 度サンプルを取り出したら元に戻さない）であっても，$p(x)$ はほとんど変化しないし，1 枚目に何がでるかという場合を総当りして合計すれば，2 枚目に何が出るかという確率も厳密に $p(x)$ となる）。このように，無作為に選び出されるカードは，すべて母集団における相対度数によって与えられる確率で，母集団の値のどれかをとりうる，**確率変数**である。

サイズ 10 の標本の標本平均 $\bar{X}$ は，$\bar{X} = \frac{1}{10}(X_1 + X_2 + ... + X_{10})$ なので，その期待値は，

$$E(\bar{X}) = \frac{1}{10}(E(X_1) + E(X_2) + ... + E(X_{10})) = \frac{1}{10}(\mu + \mu + ...\mu) = \frac{1}{10} \cdot 10 \cdot \mu = \mu$$

となって母平均に一致する。標本平均の分散は

$$V(\bar{X}) = V(\{\frac{1}{10}(X_1 + X_2 + ...X_{10})\})$$

$$= (\frac{1}{10})^2\{V(X_1) + V(X_2) + ... + V(X_{10})\}$$

$$= (\frac{1}{10})^2(\sigma^2 + \sigma^2 + ... + \sigma^2) = (\frac{1}{10})^2 \cdot 10 \cdot \sigma^2$$

$$= \frac{\sigma^2}{10}$$

したがって，標本平均の標準偏差は，$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{10}}$ となる。これは，標本から推定される母集団の平均値の信頼性を示す値なので，平均の標準誤差と呼ばれる。**標本の標準偏差が生データのばらつきを示す値であってサン**

[*1] こんなにきれいな正規分布に近い表が現実のデータとして存在するはずがなく，実は，これは R の正規乱数を使って得た人工データである。母平均 168，母標準偏差 5 の正規乱数を 100 万個発生させるためには，`X <- rnorm(1000000,168,5)` とすればよい。実行するたびに異なるデータになってしまうのを避けるためには，このコマンドを実行する前に乱数発生アルゴリズムとその初期値を `RNGkind("Mersenne-Twister",normal.kind="Inversion"); set.seed(1)` などとして固定すればよい。

プルサイズに依存しないのに対し，標準誤差は平均値の信頼性を示す値なので，一般にサンプルサイズを大きくすれば小さくできる値であるため，標準誤差を示す時はサンプルサイズも明記しなくてはならない。

一方，母集団の分散の推定値としての標本の不偏分散は，

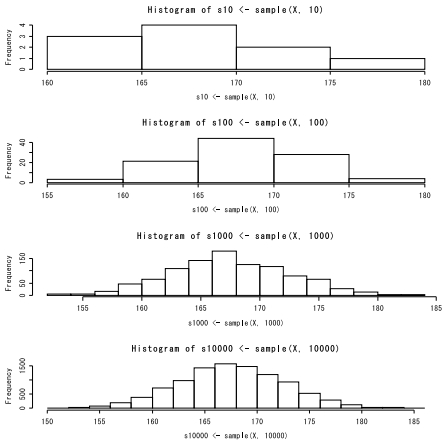
$$V(X) = \frac{1}{9}\sum_{i=1}^{10}(X_i - \bar{X})^2$$

である。なぜ $\frac{1}{10}$ でなくて $\frac{1}{9}$ かというと，$\bar{X}$ は 10 個の X_i からの計算値なので，偏差の情報は見かけ上 10 個あるようにみえるが，9 個分しかないからである（10 個の値の平均と 9 個の値が決まれば 10 個目の値はその計算値として示せてしまう）。厳密な証明はやや高度なので省略するが（参考書としては，竹村彰通『現代数理統計学』創文社をお薦めする），以下，数値シミュレーションで試してみよう。

R で標本抽出をする関数は，sample() である。replace=オプションで復元，非復元を選択できる（デフォルトは非復元）。sample() 関数は乱数を使って標本抽出しているので，乱数を初期化して揃えない限り，実行するたびに結果は変わる。

it04-1-2006.R

```
HT <- 144:191
NUM <- c(2,3,6,17,19,56,125,219,463,915,1609,2649,4550,7214,11005,16081,
 22098,29903,39048,48312,57703,66639,73332,78051,79829,77866,73767,66321,
 57993,48410,39081,29967,22055,15810,10875,7309,4596,2726,1519,939,462,
 224,128,50,31,14,5,4)
X <- rep(HT,NUM)
RNGkind("Mersenne-Twister",normal.kind="Inversion")
set.seed(1)
layout(c(1,2,3,4))
hist(s10 <- sample(X,10))
hist(s100 <- sample(X,100))
hist(s1000 <- sample(X,1000))
hist(s10000 <- sample(X,10000))
print(c(mean(s10),mean(s100),mean(s1000),mean(s10000)))
print(c(sd(s10),sd(s100),sd(s1000),sd(s10000)))
```



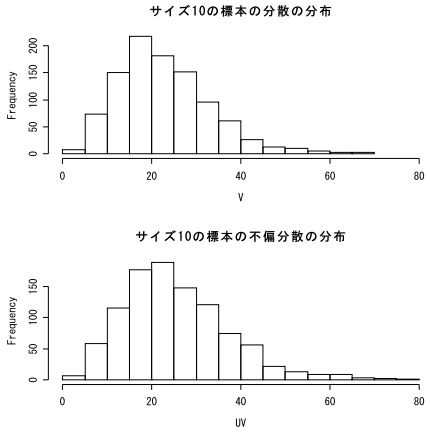
サンプルサイズが大きくなるにつれて，標本分布は母集団の分布に近づく（平均値からみても，不偏標準偏差からみても）。もっとも，ある程度大きなサンプルになると，それほどの差は感じられない。ここで標本の

分散でなくて不偏分散の方が母集団の分散に近いことをサンプルサイズが 10 の場合について確かめるには次のようにする。なお，これまで何度か出てきたが，for ( ) { }という命令は，( ) 内の条件が満たされている間，{ }内の命令を繰り返せという意味の制御文である。

variance.R

```
HT <- 144:191
NUM <- c(2,3,6,17,19,56,125,219,463,915,1609,2649,4550,7214,11005,16081,
 22098,29903,39048,48312,57703,66639,73332,78051,79829,77866,73767,66321,
 57993,48410,39081,29967,22055,15810,10875,7309,4596,2726,1519,939,462,
 224,128,50,31,14,5,4)
X <- rep(HT,NUM)
RNGkind("Mersenne-Twister",normal.kind="Inversion")
set.seed(1)
V <- rep(0,1000)
UV <- rep(0,1000)
for (i in 1:1000) {
 s10 <- sample(X,10)
 m10 <- mean(s10)
 V[i] <- sum((s10-m10)^2)/10
 UV[i] <- sum((s10-m10)^2)/9
}
layout(c(1,2))
hist(V,main="サイズ 10 の標本の分散の分布",xlim=c(0,80))
hist(UV,main="サイズ 10 の標本の不偏分散の分布",xlim=c(0,80))
print(c(mean(V),mean(UV)))
```

この結果をみると，標本の分散も不偏分散も右裾を引いた分布になっていて，分布の位置がずれている。標本の分散の期待値よりも不偏分散の平均（つまり期待値）の方が母集団の分散に近いことが確認できる。



中心極限定理

標本平均の分布については，もっと多くのことがいえる。実行するたびに標本抽出されるサンプルは異なるのだが，その平均は一定の法則に従うことが知られている。つまり，ほとんどの任意の母集団から抽出された無作為標本の平均 $\bar{X}$ の分布は，標本の大きさ n が増大するにつれて，平均 μ，標準偏差 $\sigma/\sqrt{n}$ の正規分布に

近づく[*2]これを中心極限定理という。R で試してみよう。次のプログラムは 0 から 100 まで一様分布する乱数 5000 個と平均 100，標準偏差 10 の正規分布に従う乱数 5000 個をあわせた母集団（元の分布）から，標本の大きさ（サンプルサイズ）を増やしながら標本抽出するものである。

it04-2-2006.R

```
X <- c(runif(5000,0,100),rnorm(5000,100,10))
tsd <- function(X) { sqrt(var(X)*(length(X)-1)/length(X)) }
layout(matrix(c(1,3,2,4),2,2))
hist(X,xlim=c(0,140),freq=F,main="元の分布")
Z5 <- rep(0,1000)
for (i in 1:1000) { Z5[i] <- mean(sample(X,5)) }
hist(Z5,xlim=c(0,140),freq=F)
Y2 <- dnorm(0:140,mean(X),tsd(X)/sqrt(5))
lines(0:140,Y2,col="red")
Z30 <- rep(0,1000)
for (i in 1:1000) { Z30[i] <- mean(sample(X,30)) }
hist(Z30,xlim=c(0,140),freq=F)
Y3 <- dnorm(0:140,mean(X),tsd(X)/sqrt(30))
lines(0:140,Y3,col="red")
Z200 <- rep(0,1000)
for (i in 1:1000) { Z200[i] <- mean(sample(X,200)) }
hist(Z200,xlim=c(0,140),freq=F)
Y4 <- dnorm(0:140,mean(X),tsd(X)/sqrt(200))
lines(0:140,Y4,col="red")
```

サンプルサイズが大きくなるほど，正規分布に近づくと同時に，標本平均の標準偏差（つまり標準誤差）が小さくなっていき，ヒストグラムの幅が狭くなって，理論分布（赤い線）に近づくのが一目瞭然である。

前回やったように，生データのばらつきを示す指標として標準偏差が適切なのは，生データの分布が正規分布に近い場合に限られるが，中心極限定理から考えると，平均の信頼性を示す値としての標準誤差は，データの分布によらず，ある程度サンプルサイズが大きければ常に適切な指標となる。

信頼区間

標本平均の期待値は，母平均に一致するので，標本平均は，母平均のよい点推定量になっている。しかし，必ず一致するとは限らない。そこで，それがどれくらい確からしい推定かということを示すことを考える。通常，標本から計算される，ある区間の中に，正しく μ を含む割合が 95%であるような区間を推定したい（つまり，95%の信頼度をもって区間推定をしたい）と考える。

標本が大きい場合は，$\bar{X}$ の正規分布において，ちょうど 95%の確率を囲むような最短の範囲を選択すればよい。そのために，正規分布の両側の裾 2.5%分を除く中央部分をとることができる。この部分は，母平均 μ を中心として，大きい方と小さい方に，$\bar{X}$ の標準偏差（つまり標準誤差）の 1.96 倍（1.96 は標準正規分布の 97.5%点である）だけ動かした範囲を含む。つまり，

$$Pr(\mu - 1.96\frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

[*2] μ は母平均，σ は母集団の標準偏差である。

である。この式を変形すると，

$$Pr(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

となる。σ は未知なので，通常，標本の不偏標準偏差 $sd(X)$ を用いる。つまり，μ の 95%信頼区間は，$\bar{X} - 1.96sd(X)/\sqrt{n}$ から $\bar{X} + 1.96sd(X)/\sqrt{n}$ までとなる。

標本が小さい場合は，未知の σ に対して $sd(X)$ を代入することが無視できない誤差の原因となる。そこで，$sd(X)$ を使って定義される t という統計量，

$$t = \frac{\bar{X} - \mu}{sd(X)/\sqrt{n}}$$

が自由度 $n-1$ の t 分布に従うことから，t 分布の 2.5%点から 97.5%点までを 95%信頼区間とすることで，この誤差を回避することができる。自由度が無限大になれば t 分布は正規分布に一致するので，結局，常にこちらで計算すればよいことになる。

すなわち，標本サイズ n，標本平均 $\bar{X}$，標本の不偏標準偏差 $sd(X)$ のとき，母平均の 95%信頼区間は，

$$\bar{X} - t_{0.025}sd(X)/\sqrt{n}$$

から

$$\bar{X} + t_{0.025}sd(X)/\sqrt{n}$$

までになる。R で自由度 $n-1$ の t 分布の 97.5%点を与える関数は `qt(0.975,n-1)` なので，R のプログラム上で，例えば `it04-1-2006.R` で定義した X からのサイズ 100 のサンプル s100 から母平均の 95%信頼区間を推定するには，次のようにすればよい。

it04-3-2006.R

```
HT <- 144:191
NUM <- c(2,3,6,17,19,56,125,219,463,915,1609,2649,4550,7214,11005,16081,
 22098,29903,39048,48312,57703,66639,73332,78051,79829,77866,73767,66321,
 57993,48410,39081,29967,22055,15810,10875,7309,4596,2726,1519,939,462,
 224,128,50,31,14,5,4)
X <- rep(HT,NUM)
RNGkind("Mersenne-Twister",normal.kind="Inversion")
set.seed(1)
s100 <- sample(X,100)
barX <- mean(s100)
sdX <- sd(s100)
t975 <- qt(0.975,length(s100)-1)
rootn <- sqrt(length(s100))
print(barX - t975*sdX/rootn)
print(barX + t975*sdX/rootn)
```

この結果から，95%信頼区間を表示するときは，$[167.47, 169.19]$ のように記載するのが普通である。ただし，もちろん，R にはもっと楽にこの計算をしてくれる関数が用意されていて，標本データが付値されている数値型変数 X について，`t.test(X)` とすれば，母平均がゼロであるという帰無仮説の検定結果（検定については後日説明する）とともに，母平均の推定値と母平均の 95%信頼区間が表示される。`t.test(s100)` が表示する母平均の推定値と 95%信頼区間は，上のコードで得られる結果と一致するはずである。各自確認されたい。

自由度

不偏分散の計算のところで出てきたが，自由度について直感的に理解するためには，次のように考えることもできる。n 個の数値からなる標本には，はじめ n の自由度がある。しかし，自由度の 1 つは，$\bar{X}$ を計算するときに使われてしまい，$sd(X)$ を計算するための偏差 $X_i - \bar{X}$ に対しては $n-1$ の自由度しか残らない。逆に見れば，サイズ n の標本について，はじめの $n-1$ 個の偏差は自由だが，最後の偏差は，全偏差の合計がゼロにならねばならない（そのように平均という量はとられている）ために決定済みであり，結局自由に決められる個数は $n-1$ となる。

一般には，**データの数から，推定した母数の数を引いた値が，その統計量や分布の自由度になる。**

課題

50 から 99 までの値が 1000 ずつ，合計 50000 個の値からなる母集団があるとする。そこから 5 個のサンプルを 100 回抽出した時と 25 個のサンプルを 100 回抽出したときの標本平均の分布を図示してパワーポイントに貼り付け，2 つの分布を比較して考察したことを記入し，学生証番号と氏名を打って印刷したものに氏名を自筆して提出せよ。