1　T.W.Anderson(2003) で与えられる母固有値 λ_i の $100(1-\alpha)\%$ 信頼区間の検証

1.1　扱うデータの仮定と構成した漸近的な信頼区間の確認

以下は **Theorem 13.5.1.** と同様の定義や仮定と，その仮定の下 SUBSECTION 11.6.1. で導出された母固有値 λ_i の漸近的な $100(1-\alpha)\%$ 信頼区間である．

母集団のデータベクトルは p 次元正規分布 $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ に独立同分布に従うとして，そこから抽出される標本 $\boldsymbol{x}_1,...,\boldsymbol{x}_N$ の標本共分散行列は $\boldsymbol{S}=\frac{1}{n}\sum_{\alpha=1}^{N}(\boldsymbol{x}_\alpha-\bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha-\bar{\boldsymbol{x}})'$ である．ただし，$n=N-1$ である．

母共分散行列 $\boldsymbol{\Sigma}$ の固有値を $\lambda_1>\cdots>\lambda_p$ とする．

標本共分散行列 $\boldsymbol{S}$ の固有値を $l_1>\cdots>l_p$ とする．

母固有値 λ_i についての有意水準 α の仮説検定において，帰無仮説 $\lambda_i=\lambda_i^0$ を考えるとき，検定統計量を $\sqrt{\frac{n}{2}}\frac{l_i-\lambda_i^0}{\lambda_i^0}$ と作る．

この検定統計量は漸近的に標準正規分布 $N(0,1)$ に従う．

$\sqrt{\frac{n}{2}}\frac{l_i-\lambda_i^0}{\lambda_i^0}$ は漸近的に $N(0,1)$ に従うため，帰無仮説 $H:\lambda_i=\lambda_i^0$ における有意水準 α の両側検定の（漸近的な）受容域は

(4)　$$-z(\alpha)\le\sqrt{\frac{n}{2}}\frac{l_i-\lambda_i^0}{\lambda_i^0}\le z(\alpha),$$

である．ただし，$z(\alpha)$ は $N(0,1)$ の上側 $100(\alpha/2)\%$ 点，すなわち $P_Z^H\{Z>z(\alpha)\}=\alpha/2$ を満たす点である（ただし $Z\sim N(0,1)$）．[*1]

区間 **(4)** は，信頼係数 $1-\alpha$ を持つ λ_i の信頼区間を与える（漸近分布により区間が求められるから，正確には近似的な信頼区間である）．すなわち，**(4)** を λ_i の不等式になるように簡単な式変形を行うと，以下のような信頼係数 $1-\alpha$ の λ_i の信頼区間を与える．

(5)　$$\frac{l_i}{1+\sqrt{2/n}z(\alpha)}\le\lambda_i\le\frac{l_i}{1-\sqrt{2/n}z(\alpha)}.$$

実現値 l_i について，区間 **(5)** は λ_i の $100(1-\alpha)\%$ 信頼区間である．

1.2　検証の目的と方法

T.W.Anderson で与えられた信頼係数 $1-\alpha$ を持つ λ_i の信頼区間

(5)　$$\frac{l_i}{1+\sqrt{2/n}z(\alpha)}\le\lambda_i\le\frac{l_i}{1-\sqrt{2/n}z(\alpha)},$$

は標本数 N について漸近的に与えられるものだった．

従って，（母集団のデータ数に対して）どの程度標本数 N があれば **(5)** が正しく λ_i の $100(1-\alpha)\%$ 信頼区間を与えるのか確認することがこれから行う検証の目的である．

方法は，まず 5 次元正規分布 $N(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)$ のパラメータである母期待値ベクトル $\boldsymbol{\mu}_0$ と母共分散行列 $\boldsymbol{\Sigma}_0$ を与える．この分布を母集団の分布と考える．そして，この母共分散行列の母固有値 λ_i^0 を計算する．

この母集団から標本を $N=2,3,...$ 個と無作為抽出していく．（1 つのみ抽出する場合は検定統計量 $\sqrt{\frac{n}{2}}\frac{l_i-\lambda_i^0}{\lambda_i^0}$ が標本固有値と母固有値によらず 0 になるため考えない．）

N 個の標本を並べた行列を用いて標本共分散行列の標本固有値 l_i を計算する．ここで，今は両側検定を考えるため検定統計量 $\sqrt{\frac{n}{2}}\frac{l_i-\lambda_i^0}{\lambda_i^0}$ の絶対値を計算する．

[*1] 一般に，確率変数 X の確率密度関数 f_X を持つ分布について，$0<\alpha<1$ に対して $\int_z^\infty f_X(x)dx=\alpha$ となる z を上側 $100\alpha\%$ 点という．

そして，その値が $100(1-\alpha)\%$ 信頼区間

(5) $\frac{l_i}{1+\sqrt{2/n}z(\alpha)} \leq \lambda_i \leq \frac{l_i}{1-\sqrt{2/n}z(\alpha)}$,

を満たすように得られているか確認する．受容域と信頼区間の関係より，$\left|\sqrt{\frac{n}{2}}\frac{l_i-\lambda_i^0}{\lambda_i^0}\right| \leq z(\alpha)$ を満たすことは，標本数 N に対して λ_i が $100(1-\alpha)\%$ 信頼区間

(5) $\frac{l_i}{1+\sqrt{2/n}z(\alpha)} \leq \lambda_i \leq \frac{l_i}{1-\sqrt{2/n}z(\alpha)}$,

に含まれていることと同値であることから，$\sqrt{\frac{n}{2}}\frac{l_i-\lambda_i^0}{\lambda_i^0}$ の絶対値が設定した有意水準 α の下での $z(\alpha)$ 以下になるのかを確認すればよい．

$100(1-\alpha)\%$ 信頼区間の定義より，不等号の判定 $\left|\sqrt{\frac{n}{2}}\frac{l_i-\lambda_i^0}{\lambda_i^0}\right| \leq z(\alpha)$ を各標本数 N で 100 回ずつ抽出し直した標本の集まりで繰り返し，母固有値 λ_i が 100 回の内 $100(1-\alpha)$ 回構成した信頼区間に含まれているかを確認すれば，**(5)** が正しく λ_i の $100(1-\alpha)\%$ 信頼区間を与えているかどうかの検証ができる．

ただし，この 100 回はこの検証の精度のようなもので，$100 * \mathrm{Trial}$（$\mathrm{Trial} = 1, 10, 100$）と変化させていく．

1.3 構成した漸近的な信頼区間の確認を行うコード

リスト 1　構成した漸近的な信頼区間 **(5)** $\frac{l_i}{1+\sqrt{2/n}z(\alpha)} \leq \lambda_i \leq \frac{l_i}{1-\sqrt{2/n}z(\alpha)}$ の確認を行う関数のプログラムコードとその実行

```
1  library(MASS) # 多変量正規分布に従う乱数を生成するmvrnorm()関数を用いるため，
2                # それが定義されているMASSライブラリを読み込む．
3
4  # T.W.Andersonで導出された検定統計量を用いて信頼区間を構成し，母固有値に対して漸近的な受容域(4)の検証を行う関数
5  Anderson <- function(A, N, L, Lam){       # 引数は有意水準A/100のA%，標本数N，標本固有値L，母固有値Lam
6
7    z <- qnorm(1 - (A / 2) / 100, 0, 1)        # 100(A/2/100)%点zの導出
8    T <- abs(sqrt((N - 1) / 2) * (L - Lam) / Lam) # T.W.Andersonで与えられる検定統計量の絶対値
9
10   return(if(T <= z) {1} else {0}) # T.W.Andersonの不等式(4)に対応している．
11                                   # 与えられた母固有値と標本固有値が(4)を満たせば1，満たさなければ0を返す．
12
13 }
14
15 # 期待値ベクトルMu，共分散行列Sigmaの多変量正規分布に従う母集団から，無作為抽出する標本の数を増やしていき，
16 # 各標本数毎に(100 * Trial)回のAnderson()関数を用いた信頼区間の検証を行う．
17 # (100 * Trial)回はその都度標本を取り直す．
18 # 信頼区間の検証とはつまり，抽出した標本の集まり毎の標本固有値で構成されるAndersonの信頼区間に母固有値が
19 # 含まれているかを確認し，さらに構成された信頼区間が母固有値の100(1−A/100)%信頼区間
20 # (Aは百分率として与えられるものとする)として成り立っているのかを確認し，
21 # 各標本数毎に構成した信頼区間に母固有値が含まれた回数をプロットする関数
22 ConfiPlot <- function(A, Mu, Sigma, Max, Trial){ # 引数は有意水準A/100のA%，母集団が従う多変量正規分布の
23                                                  # 期待値ベクトルMuと共分散行列Sigma，抽出する標本数の最大値Max，
24                                                  # 各標本数で試行する回数(100 * Trial)のTrial
25   NVec <- c(2:Max) # 抽出する標本数を並べるベクトル．
26                    # NVec[i] = 1のとき検定統計量が0になってしまうことを避ける．
27   CountsVec <- numeric(Max - 1) # 構成した信頼区間に母固有値が入った回数(実際はTrialで割る)を並べるベクトル
28
29   Lam <- eigen(Sigma)$values[1] # 母固有値
30
31   Guide <- 0 # 考えている有意水準A/100に対して，
32              # 上手く100(1−A/100)%信頼区間が構成できたときの標本数を入れるオブジェクト
33
34   for(i in 1:(Max - 1)){ # 抽出する標本数の最大値(Maxで与えられる)−1の回数だけ繰り返す．
35                          # つまり各標本数で以下を実行する．
```

```
36
37      Counts <- 0  # 各標本数において構成した信頼区間に母固有値が入った回数を入れるオブジェクト
38
39      for(j in 1:(100 * Trial)){ # (100 * Trial)の回数だけ繰り返し各標本数で信頼区間を構成する.
40                                 # Anderson()の結果がCountsに蓄積される.
41
42        Sam <- mvrnorm(NVec[i], Mu, Sigma) # 多次元正規分布に従う確率ベクトルをNVec[i]組発生
43                                           # ただしSigmaは対称行列で正定値行列
44        L <- (prcomp(Sam)$sdev[1]) ^ 2     # 標本固有値
45        Counts <- Counts + Anderson(A, NVec[i], L, Lam) # 標本数NVec[i]をj回目に抽出して計算したAnderson()が
46                                                        # j-1回目までのCountsに足される.
47
48      }
49
50      CountsVec[i] <- Counts / Trial # CountsVec[i]に標本数NVec[i]の時のCountsを代入する.
51      if(((CountsVec[i]) >= 100 - A) && (Guide == 0)){ # 満たしたい有意水準A/100に対して,
52                                                       # 100(1-A/100)%信頼区間が構成出来た時の
53                                                       # 標本数をGuideに代入する.
54          Guide <- NVec[i]
55      }
56    }
57
58    plot(NVec, CountsVec, xlim = c(0, Max), ylim = c(0, 100) # 横軸を標本数NVec, 縦軸を各標本数において構成した
59                                                             # 信頼区間に母固有値が入った回数CountsVecとしたグラフ
60         , xaxt = "n", yaxt = "n", xlab = "抽出した標本の数(NVec)"
61         , ylab = "構成した信頼区間に母固有値が入った回数/Trial(CountsVec)", pch = 1)
62    abline(h = 100 - A, col = 'red')                         # 横軸に平行で. 満たしたい(信頼係数*100)%に線を引く.
63
64    axis(side = 2, at = c(0, 50, 100), labels = c(0, 50, 100), cex.axis=0.6)
65    axis(side = 2, at = 100 - A,  labels = 100 - A, col.ticks ='red', col.axis = "red")
66
67    axis(side = 1, at = c(0, Max / 2, Max), labels = c(0, Max / 2, Max), cex.axis=0.6)
68    if (Guide != 0){           # Guideの値が標本数の最大値として与えられるMax以下であれば, 縦軸に平行で
69                               # 満たしたい有意水準A/100に対して100(1-A/100)%信頼区間を
70                               # 最初に構成出来た標本数で線を引く.
71       abline(v = Guide, col = 'black')
72       axis(side = 1, at = Guide,  labels = Guide, col.ticks ='black', col.axis = "black")
73    }
74  }
75
76
77  Mu0 <-  c(1, 2, 3, 4, 5) #母期待値ベクトル
78  Sigma0 <- rbind(          #母共分散行列, ただし正定値
79    c(1, 0, -1, 1, -1),
80    c(0, 2, 0, 1, 1),
81    c(-1, 0, 3, 0, 2),
82    c(1, 1, 0, 4, -1),
83    c(-1, 1, 2, -1, 5)
84  )
85
86  # 上記の母期待値ベクトルMu0, 母共分散行列Sigma0に従う確率ベクトルから標本を最大でMax = 100まで
87  # 抽出し, 標本数2,...,Maxそれぞれの下で(100 * Trial) = 100の回数だけAndersonを計算する.
88  # 有意水準はA/100 = 0.1 (A=10%) で, 100(1-A/100) = 90%信頼区間を構成したい.
89  ConfiPlot(10, Mu0, Sigma0, 100, 1)
```

各行について解説を行う.

1：プログラム内で多変量正規分布を生成するため, それを出力出来る関数 mvrnorm() 関数が定義されている MASS ライブラリを呼び出している.

5 – 13：SUBSECTION 11.6.1. にて, 母固有値 λ_i の漸近的な $100(1-\alpha)\%$ 信頼区間（このプログラムでは有意水準が

百分率のまま与えられることを想定しているため，プログラム内と解説では α が A/100 となっている）を与える検定統計量として導出された $\sqrt{\frac{n}{2}}\frac{l_i-\lambda_i^0}{\lambda_i^0}$ を計算し，そこで考えている標本から与えられる信頼区間が母固有値 λ_i を含むのかを確認する Anderson 関数の定義が行われている．実際に含めば 1 を，含まなければ 0 を返す Anderson() 関数である．ここでは，100(1 − A/100)% 信頼区間と有意水準 A/100 の両側検定の採択域の対応や，abs() 関数を用いるなどして立式を簡単にしている．具体的には，100(1 − A/100)% 信頼区間 **(5)** $\frac{l_i}{1+\sqrt{2/n}z(\alpha)} \le \lambda_i \le \frac{l_i}{1-\sqrt{2/n}z(\alpha)}$ ではなく，検定統計量 $\sqrt{\frac{n}{2}}\frac{l_i-\lambda_i^0}{\lambda_i^0} = \sqrt{\frac{\mathrm{N-1}}{2}}\frac{\mathrm{L-Lam}}{\mathrm{Lam}}$ による有意水準 A/100 の両側検定の採択域 **(4)** $-z(\alpha) \le \sqrt{\frac{n}{2}}\frac{l_i-\lambda_i^0}{\lambda_i^0} \le z$ を用いて確認を行っている．仮引数は有意水準 $(\alpha \times 100)$%（A）と標本数 N，標本固有値 l_i（L），母固有値 λ_i（Lam）である．このように検定統計量を Anderson() 関数として独立させたのは，検定統計量や棄却限界値の分布を変えて試行できるようにするためである．

7：A は百分率に変換された有意水準として与えられることを想定するため，それを 1 未満の確率の値（両側検定であるからここではその半分）に変換しつつ，採択域を考えるためにその値を 1 から引いている（1 − (A/2)/100）．そして，平均 μ，標準偏差 σ に従う確率変数が x 以下の値を取る確率を返す qnorm(x, μ, σ) 関数を用いて，標準正規分布で 1 − (A/2)/100 以下の値を得る確率，つまり 100(alpha/100/2)% 点を z として定義している．

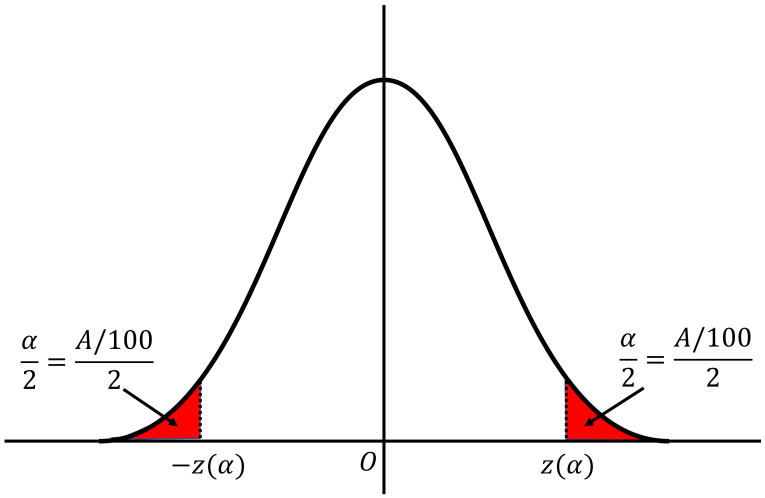

図 1　$\sqrt{\frac{n}{2}}\frac{l_i-\lambda_i}{\lambda_i}$ の漸近分布 $N(0,1)$ の密度関数

8：T を検定統計量 $\sqrt{\frac{n}{2}}\frac{l_i-\lambda_i^0}{\lambda_i^0} = \sqrt{\frac{\mathrm{N-1}}{2}}\frac{\mathrm{L-Lam}}{\mathrm{Lam}}$ の絶対値として定義している．

10：検定統計量 T が 100(A/100/2)% 点 z 以下，つまり仮定が採択されるような標本が取れたときに 1 を返し，取れなかった場合は 0 を返す．

22 − 74：母集団に対して，無作為抽出する標本の数 N を増やしていき，それぞれの標本の集まりで Anderson() 関数を用いて信頼区間の検証を行う．信頼区間の検証とはつまり，それぞれの標本の集まりの標本固有値 l_i（L）で構成される Anderson の信頼区間に母固有値 λ_i（Lamda）が含まれているかを 100∗Trial 回ずつ確認し，与えられた信頼区間が母固有値の 100(1 − A/100)% 信頼区間として成り立っているのかを確認してその結果をグラフで表す ConfiPlot() 関数である．仮引数は有意水準 α（A/100）と母集団が従う多変量正規分布の期待値ベクトル（Mu），共分散行列（Sigma），各標本数における試行回数の倍数 (Trial)，抽出する標本数の最大値 (Max) である．従って，有意水準や母集団が従う多変量正規分布のパラメータ，各標本数毎の試行回数を任意に変えて試行できる．

25：標本数 N を並べたベクトルとして，NVec = 2, 3, ..., Max を定義している．抽出する最大の標本数は仮引数 Max として受け取る．ただし，NVec[i]= 1 のとき検定統計量が母固有値と標本固有値に関係無く 0 になるため，常に信頼区間が正しいものと成ってしまう．従って，NVec[i] は 2 以上とする．

27：それぞれの標本数における検証の結果，つまり同じ標本数で 100∗Trial 回標本を抽出し，抽出する毎の Anderson() 関数の結果を足していった結果を，標本数と結果が対応するように並べたベクトルとして，長さ Max−1 のベクトルを定義している．この長さは抽出する標本数が N= 2,3,...,Max と変化するためであり，最大の個数が Max である．numeric(n) はデータ型 numeric（整数や実数）の n 個のオブジェクトを持つベクトルを生成する関数である．

29：仮引数として受け取った母共分散行列から母固有値 Lam を計算する．eigen() 関数により計算し，特にその第一成分，つまり第一主成分に対応する固有値を Lam として定義している．

31：Guide に 0 を代入．これは確認したい有意水準 A/100 を超えるような 100(1-A/100)% 信頼区間が始めて構成出来た時の標本数 N を代入するオブジェクトとして利用する．

34 – 48：この for 文は仮引数として受け取った，抽出する標本数の最大値 Max−1 の回数だけ繰り返される．各標本数 NVec[i] で Anderson() 関数による信頼区間の検証を 100∗Trial 回行い，その結果をそれぞれ Counts に足していき，Counts/Trial（ここで Trial で割るのは，試行回数を有意水準と対応づけるため最大でも 100 に収まるようにするためである）を標本数 NVec[i] に対応するベクトル CountsVec の要素 CountsVec[i] に代入する．そして，確認したい有意水準 A/100 を超えるような 100(1-A/100)% 信頼区間が構成出来た標本数 NVec[i] が確認出来れば，その標本数を Guide に代入する．

37：各標本数 NVec[i] で Anderson() 関数による信頼区間の検証を 100∗Trial 回行い，その結果を代入していくオブジェクトとして Counts を初期値 0 で定義している．

39 – 48：この for 文は 100∗Trial 回繰り返される．標本数が NVec[i] 個の時の Anderson() 関数による検証を 100∗Trial 回行い，結果を Counts に 100∗Trial 回分足していく．ここで用いられる仮引数として与えられる Trial は検証の精度に関わるオブジェクトであり，この値を大きくすれば構成する信頼区間が得たい信頼区間に収束することが確認出来る．

42：標本を並べた行列 Sam として，考える標本数 NVec[i] の数だけ母集団からベクトルを得て行列に並べている．この母集団は仮引数として受け取る期待値ベクトル Mu，共分散行列 $Sigma$ をパラメータとして持つ多変量正規分布に従い，それを mvrnorm() 関数で乱数生成している．この無作為抽出は 100∗Trial 回抽出し直されている．

44：抽出した標本を並べた行列 Sam の共分散行列の固有値を計算し，特にその第一成分，つまり第一主成分に対応する固有値を L に代入している．ここでは共分散行列を求める手間を主成分分析が行える関数 prcomp() 関数を用いることで省いている．

45：標本数 NVec[i] で抽出した標本固有値 L と母固有値 Lam を用いて Anderson() 関数で検証を行い，適切に 100(1 – A/100)% 信頼区間が与えられていれば 1 を，与えられなければ 0 を Counts に足している．

50：Counts/Trial を標本数 NVec[i] に対応するベクトル CountsVec の要素 CountsVec[i] に代入する．ここで Trial で割るのは，試行回数を有意水準と対応づけるため最大でも 100 に収まるようにするためである．例えば A= 5 であれば，この 100∗Trial 回の繰り返しが終わり CountsVec[i]≥ 95 となっていれば，その標本数 NVec[i] で 100(1 – A/100)% 信頼区間が適切に構成出来たことになる．

51 – 56：確認したい有意水準 A/100 を超えるような 100(1-A/100)% 信頼区間が構成出来た標本数 NVec[i] が確認出来れば，その標本数を Guide に代入する．つまり，CountsVec[i] が (100 – A)% を初めて超えた時，その値を Guide に代入する．ただし，T.W.Anderson で注意があった通り，今考えている漸近的な信頼区間において NVec[i] は，sqrt(2/NVec[i])·z < 1 を満たすように大きく取らなければならない．しかしここでは，NVec[i] が小さいと CountsVec[i] が常に 100 となってしまう等の不具合は起きていない．

58 – 73：横軸を NVec[i]（ただし [2, Max] の値を取る），縦軸を対応する CountsVec[i] とするグラフをプロットしている．考えている有意水準 A/100%，つまり 100(1 – A/100)% 信頼区間として達して欲しい A% も横軸に平行に赤色でラインを引いている．そして，そのラインを初めて超える標本数 CountsVec[i] で縦軸に平行に黒色でラインを引いている．ただ，この Guide に対応するラインは，Max 以下でない場合があるため，その場合は表示しないように if 文を設定している．

77：母期待値ベクトル Mu0 の定義．5 次元で定義している．

78 – 84：母共分散行列 Sigma0 の定義．5 次元で定義しており，正定値行列でなければならない．つまり全ての固有値が正の対称行列出なければならない．（これは行列が正定値がどうかの判定に利用できる．eigen() 関数を用いて固有値を確認すれば良い．）

89：ConfiPlot() 関数の実行．ただし，ConfiPlot(10, Mu0, Sigma0, 100, 1) で，有意水準が A/100 = 10/100（百分率で A = 10%）で，母集団が期待値ベクトル Mu=Mu0，共分散行列 Sigma=Sigma0 の多変量正規分布に従い，標本数 N として取る最大値が Max= 100，試行回数が 100∗Trial= 100 ∗ 1 = 100 回である．

1.4　リスト 1 の実行結果と考察

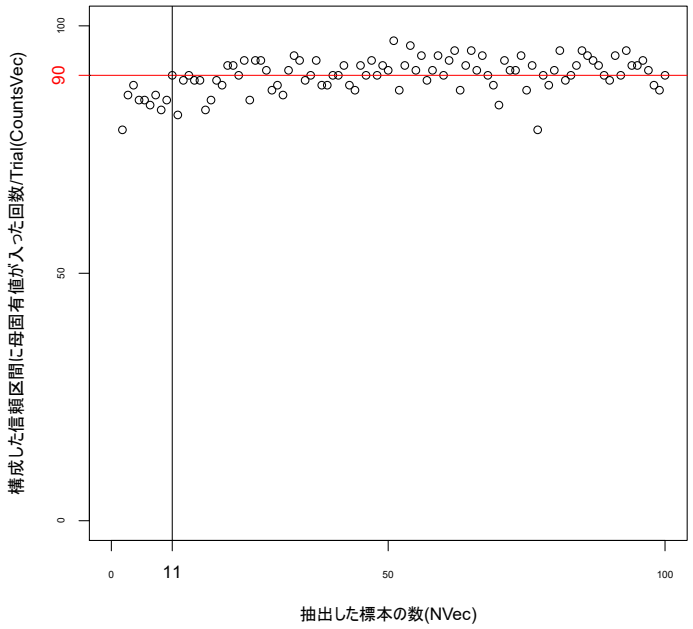

図 2　有意水準 A/100= 1.0，標本数の最大値 Max= 100，試行回数 100∗Trial= 100 のグラフ（ConfiPlot(10, Mu0, Sigma0, 100, 1) の実行結果）

図 2：ConfiPlot(10, Mu0, Sigma0, 100, 1) の実行結果，つまり有意水準が 10/100（百分率で A = 10%）で，標本数の最大値 Max= 100，試行回数 100∗Trial= 100 としている．ただし，期待値ベクトルは

$$\mathrm{Mu0} = (1,2,3,4,5)' \tag{1}$$

で共分散行列は

$$\mathrm{Sigma0} = \begin{pmatrix} 1 & 0 & -1 & 1 & -1 \\ 0 & 2 & 0 & 1 & 1 \\ -1 & 0 & 3 & 0 & 2 \\ 1 & 1 & 0 & 4 & -1 \\ -1 & 1 & 2 & -1 & 5 \end{pmatrix} \tag{2}$$

である．

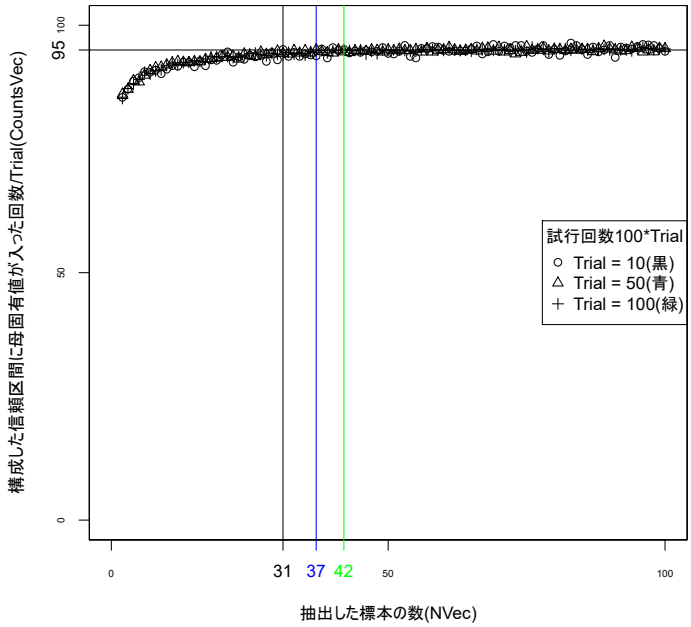

図 3　有意水準 A/100 = 0.5，標本数の最大値 Max= 100 とした上で，試行回数 100∗Trial= 1000,5000,10000 と変化させていったグラフ

図 3：先ほどと期待値ベクトルと共分散行列は同じである．有意水準は 0.5 としている．試行回数 100∗Trial= 1000,5000,10000 はそれぞれ丸，三角，十字と対応しており，最初に有意水準を満たす標本数でのラインが黒，青，緑と対応している．この結果から，より正確に信頼区間を構成しようとすればより多くの試行回数が必要であるとわかる．また，図 2 と比較すれば，各標本毎の試行回数が少ないと，初めて構成したい信頼区間を構成出来るような標本数以降の標本数でも，構成したい信頼区間が構成できていないことがある．つまり，試行回数が多くなければ検定統計量の漸近正規性を保証できない．

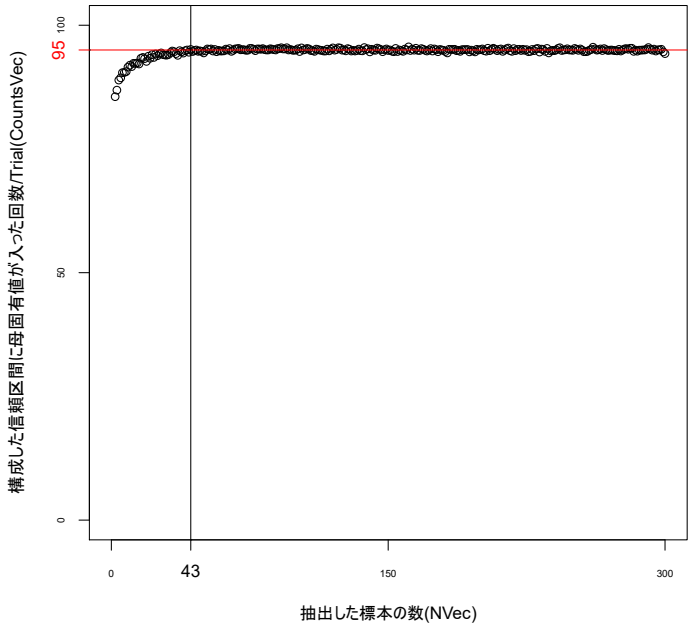

図 4　有意水準 A/100= 5.0，標本数の最大値 Max= 300，試行回数 100∗Trial= 10000 のグラフ（ConfiPlot(5, Mu0, Sigma0, 300, 100) の実行結果）

図 4：ConfiPlot(5, Mu0, Sigma0, 300, 100) の実行結果である．試行回数を増やし図 2 に比べて正確なグラフを描画すると，構成する信頼区間の信頼係数が収束しているように見える．従って，十分な標本数（今考えている母集団であれば N= 43）があれば，100(1 − A/100) = 95% 信頼区間が構成出来ることがわかった．

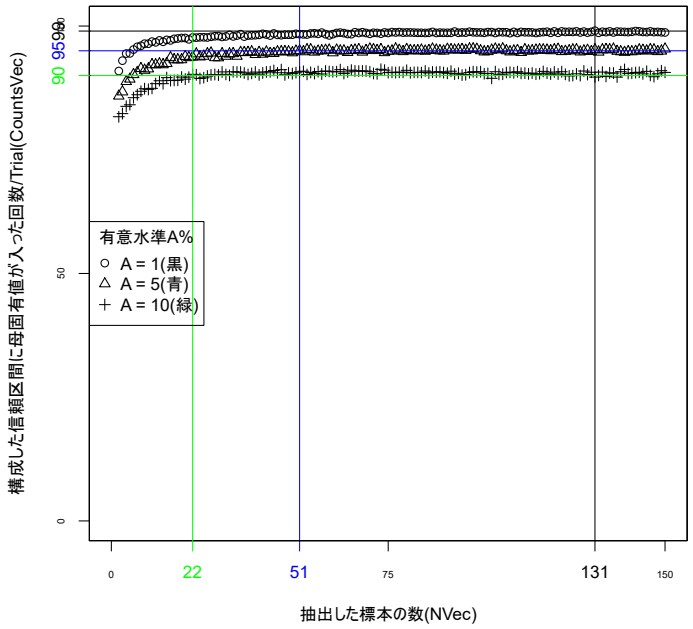

図 5　標本数の最大値 Max= 150，試行回数 100∗Trial= 100 とした上で，有意水準を A/100 = 0.1, 0.5, 1.0 と変化させていったグラフ

図 5：先ほどまでと期待値ベクトルと共分散行列は同じである．有意水準 A/100 = 0.1, 0.5, 1.0 はそれぞれ丸，三角，十字と対応しており，最初に有意水準を満たす標本数でのラインが黒，青，緑と対応している．これより，より狭い信頼区間はより少ない標本数で構成できることがわかる．

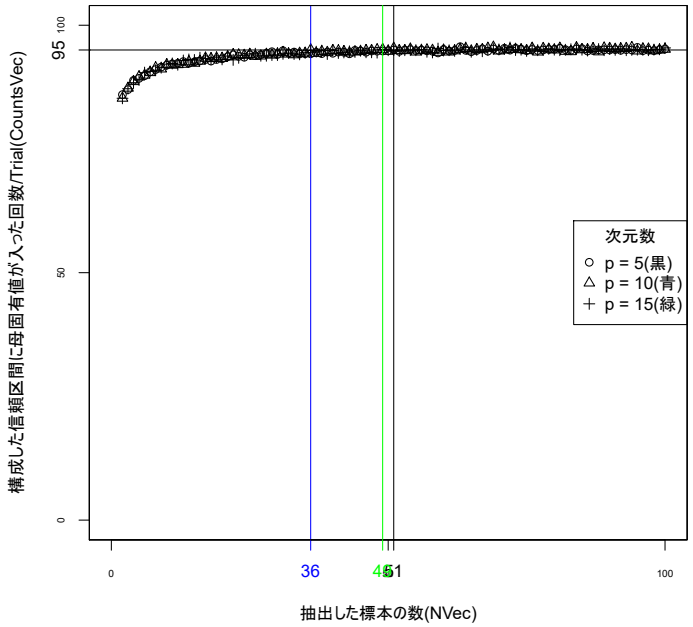

図 6　有意水準 A/100 = 0.5，標本数の最大値 Max= 100，試行回数 100∗Trial= 100 とした上で，母集団の次元数を 5,10,15 と変化させていったグラフ

図 6：母集団の次元数 5,10,15 はそれぞれ丸，三角，十字と対応しており，最初に有意水準を満たす標本数でのラインが黒，青，緑と対応している．期待値ベクトルや共分散行列の要素にも依ると思うが，データの次元数というよりも共分散行列の固有値（ここでは第一主成分の固有値）の値に依るから，次元数が大きいから標本数も必要だとは言えないことがわかる．もっと極端な例を見たい．

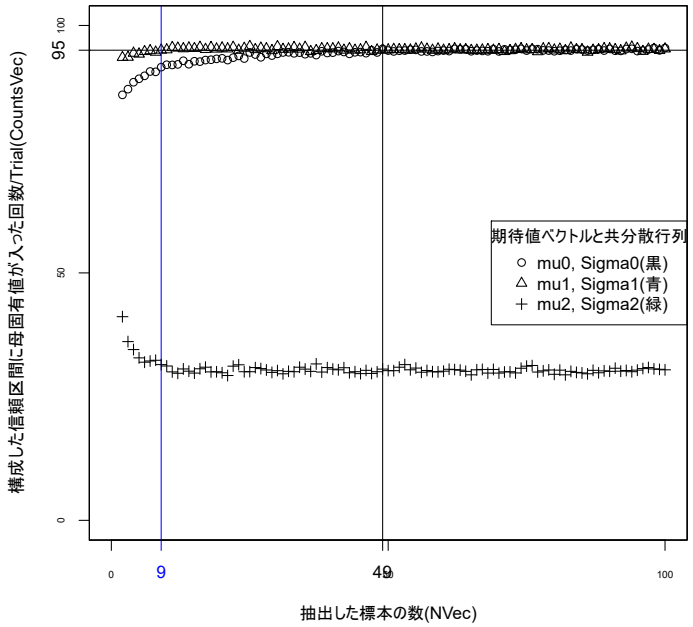

図 7　有意水準 0.5，標本数の最大値 Max= 100 の下で，母集団である 5 次元正規分布の持つパラメータの値を変化させていったグラフ

図 7：母集団である 5 次元正規分布の持つパラメータの値を変化させている．それぞれ丸，三角，十字と対応しており，

最初に有意水準を満たす標本数でのラインが黒，青，緑と対応している（ただし緑は無い）．丸 (mu0,Sigma0) はこれまでと同じパラメータで，三角 (mu1,Sigma1) は少し値を変えただけのパラメータである．十字 (mu2, Sigma2) は

$$\mathrm{Mu2} = (0,0,0,0,0)' \tag{3}$$

で共分散行列は

$$\mathrm{Sigma2} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{4}$$

としている．

データの無相関が信頼区間の構成に影響があるとわかる．

1.5 実行結果と考察による疑問

ここで構成した信頼区間が漸近的なものであったことから当然ではあるが，標本数と構成できる信頼区間の精度が関係することがわかった (図 5)．つまりは，標本数を増やせば構成できる信頼区間の精度も上がるということがわかった．これより，検定統計量を比較するときは母集団の数に対して使う標本数を事前に決めておく必要がある（これは当たり前で，結果から特別わかることとは言えない）．標本数を多く用いれば小さい p-値を得られる（図 (6) から，緑→青→黒と p-値として考える事象は小さくなるから，それを得る確率が小さくなることは想像できる）ため，p-値のみの議論を行うのであればこのような設定が必要であり，もしくは仮説検定において検定力の導出が必要である．

母数のデータ数や次元数で処理の時間が増えることがわかったが，オーダーはどのように計算するのか．次元数や標本数，試行回数によって処理時間が大きく変わっていた．

1.6 作成した検定統計量が標準正規分布に従うことを確認するコード

リスト 2　作成した検定統計量 $\sqrt{\frac{n}{2}}\frac{l_i-\lambda_i}{\lambda_i}$ が漸近的に標準正規分布に従うことを確認するコード

```
1  library(MASS)
2
3  T_Anderson <- function(N, L, Lam){
4
5   return (sqrt((N - 1) / 2) * (L - Lam) / Lam)
6
7  }
8
9  AndersonHist <- function(Mu, Sigma, N, Trial){
10
11   Lam <- eigen(Sigma)$values[1] # 母固有値
12
13   L <- numeric(100 * Trial) # 標本を抽出する度にその固有値を入れるためのベクトル
14   Anderson <- numeric(100 * Trial) # (100 * Trial)の回数だけAnderson()の結果を得て,
15                                    # それを入れるベクトル
16
17     for(i in 1:(100 * Trial)){ # (100 * Trial)の回数だけAnderson()の結果を得る.
18                                # その結果をAndersonに蓄積する.
19
20       Sam <- mvrnorm(N, Mu, Sigma) #多次元正規分布に従う独立な確率ベクトルをN組発生．ただしSigmaは対称行列で正定値行列
21       L[i] <- (prcomp(Sam)$sdev[1]) ^ 2 # 標本固有値
22       Anderson[i] <- T_Anderson(N, L[i], Lam) # 標本数Nをi回目に抽出して計算したT_Anderson()を
23                                              # Anderson[i]に代入する.
24
```

```
25      }
26
27    df <- data.frame(Anderson)
28
29    library(ggplot2)
30    ggplot(data = df, aes(x = Anderson)) +
31      geom_histogram()
32  }
33
34  Mu0 <-  c(1, 2, 3, 4, 5) #母期待値ベクトル
35  Sigma0 <- rbind( #母共分散行列, ただし正定値
36    c(1, 0, -1, 1, -1),
37    c(0, 2, 0, 1, 1),
38    c(-1, 0, 3, 0, 2),
39    c(1, 1, 0, 4, -1),
40    c(-1, 1, 2, -1, 5)
41  )
42  AndersonHist(Mu0, Sigma0, 100, 1)
```

先ほどのリスト 1 とほぼ同様である.

T_Anderson() 関数で検定統計量 $\sqrt{\frac{n}{2}}\frac{l_i-\lambda_i}{\lambda_i}$ の値を 100∗Trial 回導出し，その値全ての頻度でヒストグラムを描画している.

ggplot() 関数を用いている.

描画されるヒストグラムが標準正規分布のようになれば良い.

1.7　リスト 2 の実行結果と考察

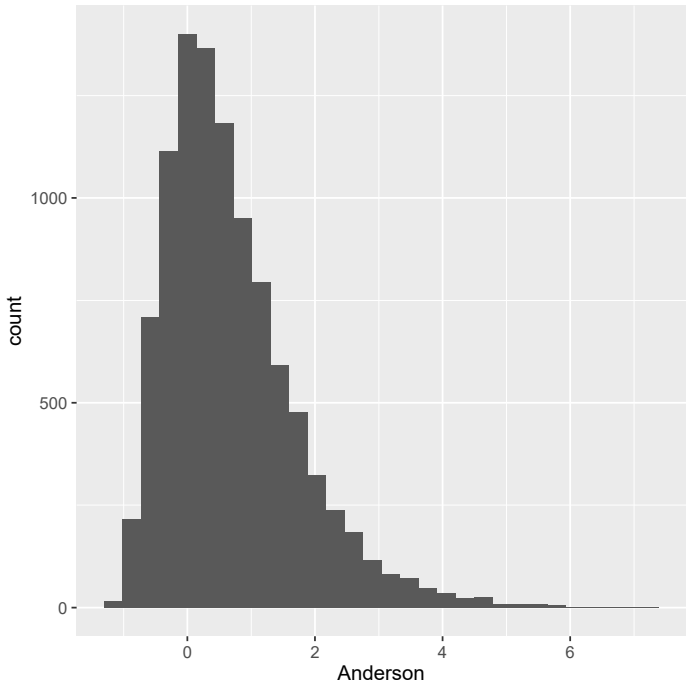

図 8　標本数の最大値 Max= 5，試行回数 100∗Trial= 10000 とした検定統計量のヒストグラム

図 8：少ない標本数でも，試行回数を増やせば単峰性は持つことがわかる.

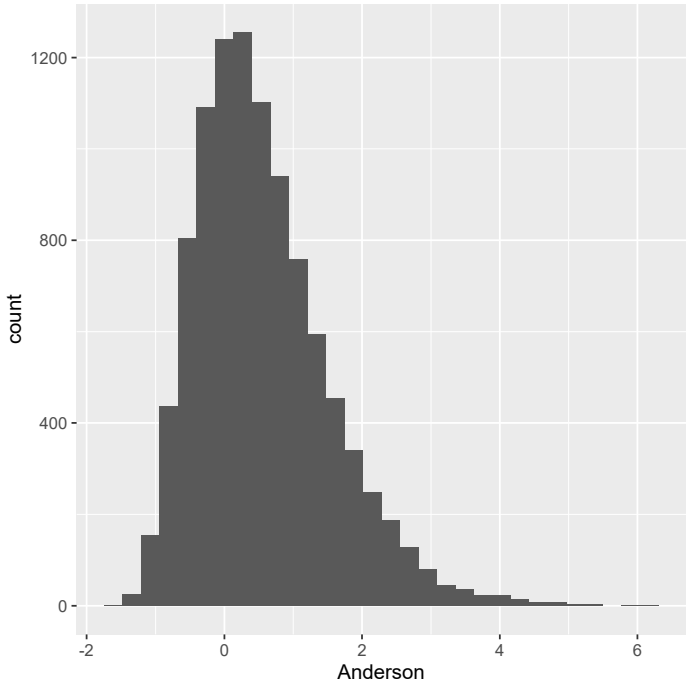

図 9　標本数の最大値 Max= 10，試行回数 100∗Trial= 10000 とした検定統計量のヒストグラム

図 9：単峰型のグラフであり，最大に取れる標本数が増えたためは Anderson が 0 を取る時の頻度が 1000 から 1200 以上と増加している．

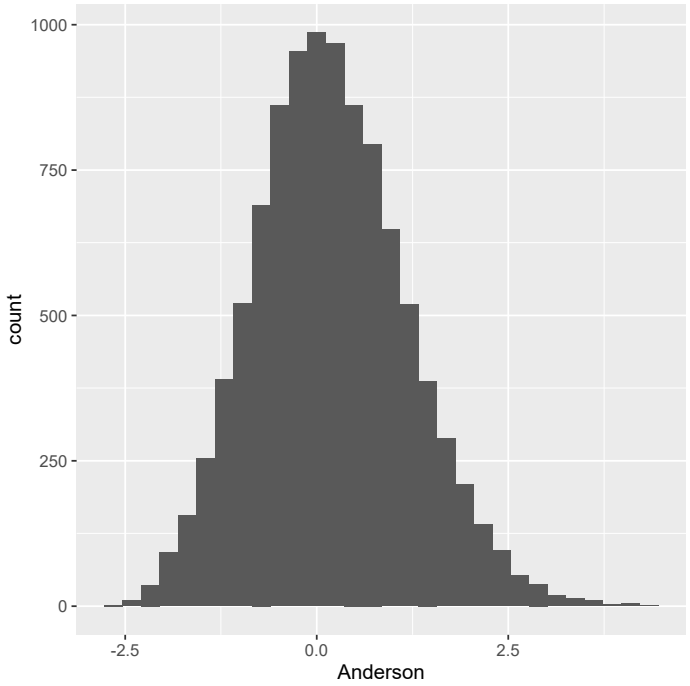

図 10　標本数の最大値 Max= 100，試行回数 100∗Trial= 10000 とした検定統計量のヒストグラム

図 10：単峰型のグラフに見えるだけでなく，平均 0 や標準偏差 1 標準正規分布に近づいて見える．

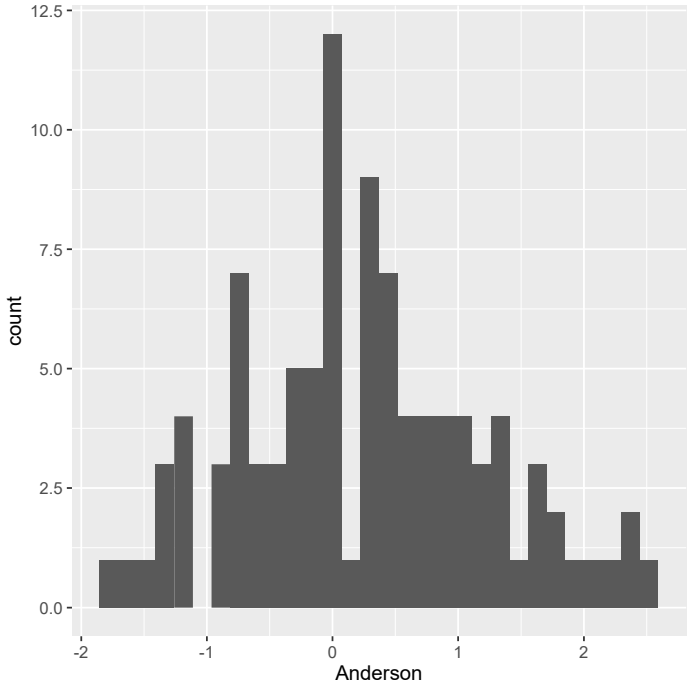

図 11　標本数の最大値 Max= 100，試行回数 100∗Trial= 100 とした検定統計量のヒストグラム

図 15：標本数の最大値と試行回数ともに少ないが，単峰型のグラフには見える．

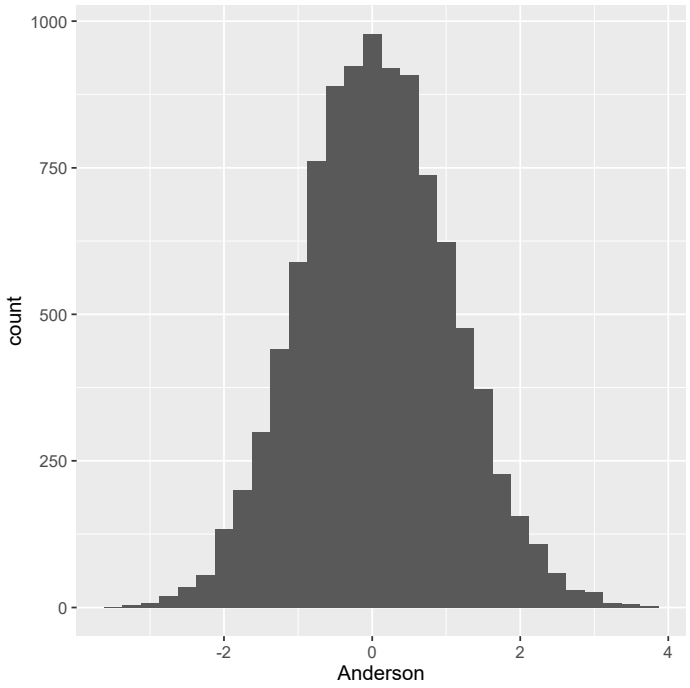

図 12　標本数の最大値 Max= 1000，試行回数 100∗Trial= 10000 とした検定統計量のヒストグラム

図 12：標準正規分布のグラフに見える．

1.8 母集団のデータ数を固定した上で構成した漸近的な信頼区間の確認を行うコード

リスト 3 構成した漸近的な信頼区間 **(5)** $\frac{l_i}{1+\sqrt{2/nz(\alpha)}} \leq \lambda_i \leq \frac{l_i}{1-\sqrt{2/nz(\alpha)}}$ の確認を行う関数のプログラムコードとその実行

```
# T.W. Andersonで導出された，漸近的な受容域(4)を与える検定統計量を用いた信頼区間の検証を行う関数
Anderson <- function(alpha, SAMPLE_NUM, SAMPLE_EV, POP_EV){

  level <- 1 - (alpha / 2) / 100
  z <- qnorm(level, 0, 1)
  T_Anderson <- abs(sqrt((SAMPLE_NUM - 1) / 2) * (SAMPLE_EV - POP_EV) / POP_EV)

 return(if(T_Anderson <= z) {1} else {0})

}

# 母集団のデータ数に対して，無作為抽出する標本の数を増やしていき，それぞれの標本の集まりで
# Anderson()関数を用いて信頼区間の検証を行う．そして，信頼区間の検証とはつまり，
# それぞれの標本の集まりの標本固有値で構成されるAndersonの信頼区間に母固有値が含まれているか
# 確認し，与えられた信頼区間が母固有値の100(1−alpha/100)%信頼区間として成り立っているのかを確認して
# 結果をグラフで表す関数
ConfiPlot <- function(alpha, POP_NUM, mu, sd, p){

  SAMPLE_NUM <- seq(5, POP_NUM, by = 5)
  Percents <- numeric(length(SAMPLE_NUM))

  POP <- matrix(rnorm(POP_NUM * p, mu, sd), ncol = p)
  POP_EV <- (prcomp(POP)$sdev[1]) ^ 2

  decide <- 0

  for(i in 1:length(SAMPLE_NUM)){

    Percent <- 0

    for(j in 1:100){
    SAMPLE <- POP[sample(1:POP_NUM, SAMPLE_NUM[i]), ]
    SAMPLE_EV <- (prcomp(SAMPLE)$sdev[1]) ^ 2
    Percent <- Percent + Anderson(alpha, SAMPLE_NUM[i], SAMPLE_EV, POP_EV)
    }

    Percents[i] <- Percent
    if((Percents[i] >= 100 - alpha) && (decide == 0))
      decide <- SAMPLE_NUM[i]
  }

  plot(SAMPLE_NUM, Percents, xlim = c(0, POP_NUM), ylim = c(0, 100)
       , xaxt = "n", yaxt = "n", xlab = "抽出した標本の数(SAMPLE_NUM)"
       , ylab = "構成した信頼区間に母固有値が入った回数(Percents)", pch = 1)
  abline(h = 100 - alpha, col = 'red')
  abline(v = decide, col = 'black')

  axis(side = 2, at = c(0, 50, 100), labels = c(0, 50, 100), cex.axis=0.6)
  axis(side = 2, at = 100 - alpha,  labels = 100 - alpha, col.ticks ='red', col.axis = "red")

  axis(side = 1, at = c(0, POP_NUM / 2, POP_NUM), labels = c(0, POP_NUM / 2, POP_NUM), cex.axis=0.6)
  axis(side = 1, at = decide,  labels = decide, col.ticks ='black', col.axis = "black")

}

```

```
56 ConfiPlot(5, 1000, 5, 10, 5)
```

1.9 リスト 3 の実行結果と考察

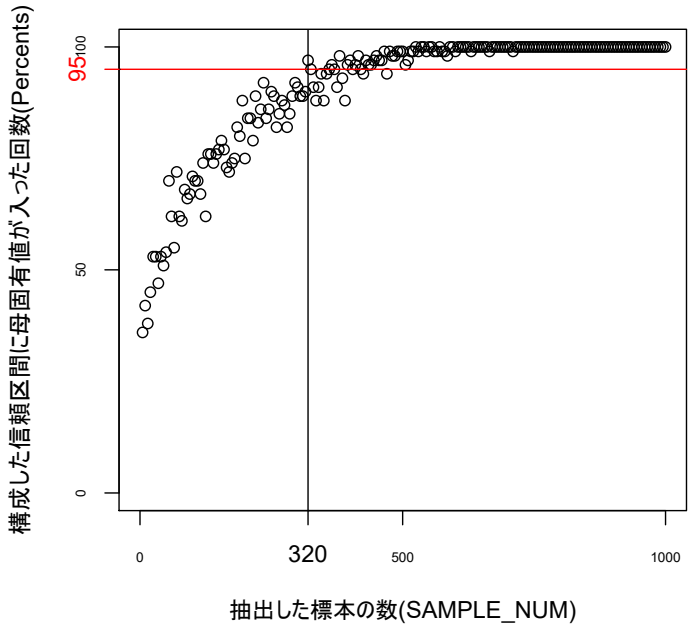

図 13　有意水準 0.5，母集団のデータ数 1000，次元数 5 のグラフ（ConfiPlot(5, 1000, 5, 10, 5) の実行結果）

1.10 母集団のデータ数を固定した上で作成した検定統計量が標準正規分布に従うことを確認するコード

リスト 4　作成した検定統計量 $\sqrt{\frac{n}{2}}\frac{l_i-\lambda_i}{\lambda_i}$ が標準正規分布に従うことを確認するコード

```
1  T_Anderson <- function(SAMPLE_NUM, SAMPLE_EV, POP_EV){
2
3   return (sqrt((SAMPLE_NUM - 1) / 2) * (SAMPLE_EV - POP_EV) / POP_EV)
4
5  }
6
7  AndersonHist <- function(POP_NUM, SAMPLE_NUM, mu, sd, p){
8
9    POP <- matrix(rnorm(POP_NUM * p, mu, sd), ncol = p)
10   POP_EV <- (prcomp(POP)$sdev[1]) ^ 2
11
12   Anderson <- numeric(100)
13
14   for(i in 1:100){
15     SAMPLE <- POP[sample(1:POP_NUM, SAMPLE_NUM), ]
16     SAMPLE_EV <- (prcomp(SAMPLE)$sdev[1]) ^ 2
17     Anderson[i] <- T_Anderson(SAMPLE_NUM, SAMPLE_EV, POP_EV)
18   }
19
20   df <- data.frame(Anderson)
21
22   library(ggplot2)
23   g <- ggplot(data = df, aes(x = Anderson))
```

```
  g <- g + geom_histogram()
  plot(g)
}

AndersonHist(1000, 500, 5, 10, 5)
```

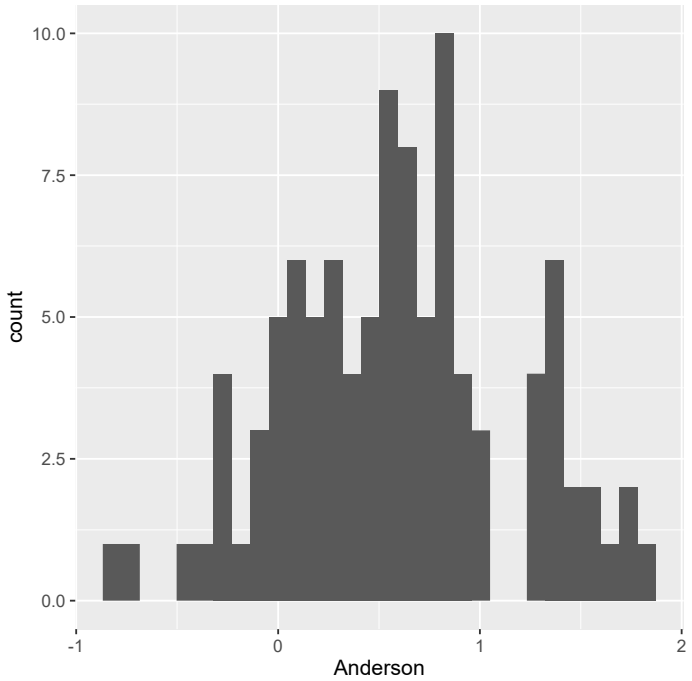

図 14　母集団のデータ数 POP_NUM = 1000，標本として抽出するデータ数 SAMPLE_NUM = 500 とした検定統計量のヒストグラム

1.11　異なる母平均と母共分散行列を持つ母集団から標本を抽出して検定統計量を計算し，それぞれのヒストグラムを同じグラフにプロットするコード

リスト 5　異なる母平均と母共分散行列を持つ母集団から標本を抽出して検定統計量を計算し，それぞれのヒストグラムを同じグラフにプロットするコード

```
library(ggplot2)
library(MASS)

T_Anderson <- function(N, L, Lam){

 return (sqrt((N - 1) / 2) * (L - Lam) / Lam)

}

AndersonHist <- function(Mu, Sigma, N, Trial){

  Lam <- eigen(Sigma)$values[1] # 母固有値

  L <- numeric(100 * Trial) # 標本を抽出する度にその固有値を入れるためのベクトル
  values <- numeric(100 * Trial) # (100 * Trial)の回数だけAnderson()の結果を得て,
                                 # それを入れるベクトル

```

```
    for(i in 1:(100 * Trial)){ # (100 * Trial)の回数だけAnderson()の結果を得る.
                               # その結果をAndersonに蓄積する.

      Sam <- mvrnorm(N, Mu, Sigma) #多次元正規分布に従う独立な確率ベクトルをN組発生. ただしSigmaは対称行列で正定値行列
      L[i] <- (prcomp(Sam)$sdev[1]) ^ 2 # 標本固有値
      values[i] <- T_Anderson(N, L[i], Lam) # 標本数Nをi回目に抽出して計算したT_Anderson()を
                                            # Anderson[i]に代入する.

    }

  df <- data.frame(values)
  return(df)
}

Mu0 <- c(1, 2, 3, 4, 5) #母期待値ベクトル
Sigma0 <- rbind( #母共分散行列, 独立な方が標本数が必要
  c(1, 0, -1, 1, -1),
  c(0, 2, 0, 1, 1),
  c(-1, 0, 3, 0, 2),
  c(1, 1, 0, 4, -1),
  c(-1, 1, 2, -1, 5)
)

Mu1 <- c(-1, -90, -3, -6, -52) #母期待値ベクトル
Sigma1 <- rbind( #母共分散行列, 独立な方が標本数が必要
  c(521, 0, -1, 1, -1),
  c(0, 235, 0, 1, 1),
  c(-1, 0, 352, 0, 2),
  c(1, 1, 0, 433, -1),
  c(-1, 1, 2, -1, 5523)
)

Mu2 <- c(0, 0, 0, 0, 0) #母期待値ベクトル
Sigma2 <- rbind( #母共分散行列, 独立な方が標本数が必要
  c(1, 0, 0, 0, 0),
  c(0, 1, 0, 0, 0),
  c(0, 0, 1, 0, 0),
  c(0, 0, 0, 1, 0),
  c(0, 0, 0, 0, 1)
)

Trials <- c(10, 10, 10)

An_a <- AndersonHist(Mu0, Sigma0, 100, Trials[1])
An_b <- AndersonHist(Mu1, Sigma1, 100, Trials[2])
An_c <- AndersonHist(Mu2, Sigma2, 100, Trials[3])

df <- rbind(An_a, An_b, An_c)
trt <- factor(rep(c("Anderson_a", "Anderson_b", "Anderson_c")
                  , Trials * 100))
df_A <- cbind(df, trt)
colnames(df_A) <- c("values", "type")

g <- ggplot(data = df_A, aes(x = values, fill = type))
g <- g + geom_histogram(position="identity", alpha = 0.8, binwidth = 0.1)
plot(g)
```

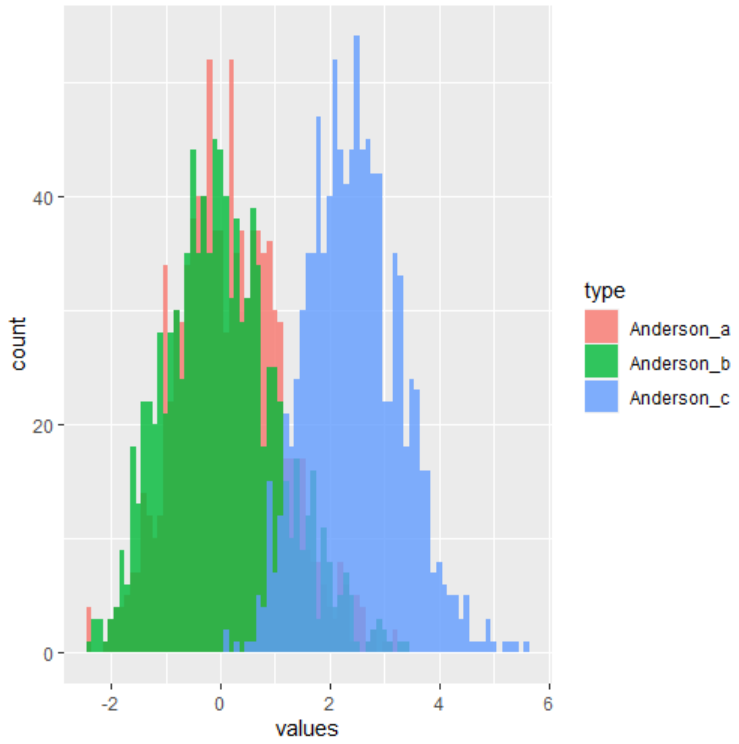

図 15　異なる母平均と母共分散行列を持つ母集団から標本を抽出して検定統計量を計算しヒストグラム