統計学の基本を理解しよう

2023.10

１　はじめに

（１）なぜデータ分析が必要なのか。　→　人間は神様ではないから。

（２）デカルト平面では、点（幾何学）を、XY 座標（代数学）で表すことができる。三次元空間の場合は XYZ 座標、四次元空間以上は図で表すことができないが、代数で考えることができる。

（３）上達の秘訣

- ・具体的な目標を持つ・・・例．不動産価格の回帰分析をしたい。
- ・具体的な疑問を持ち、それについて考える。
  - 例.「平均と分散はなぜ重要なのか？」、「不偏分散はなぜｎ－１で割るのか？」、「推定・検定は何を推定・検定するのか？」、「ｔ検定、χ^2検定はそれぞれどういったケースで使うのか？」
- ・数学から逃げない。・・・私も、「log って何だったっけ？」から学び直し。
- ・ｎ次方程式の微分ができて、Σ（シグマ）記号に対する拒否反応をなくせば何とかなる。
- ・数式を見て、どれが変数で、どれが定数なのかを意識する（例. $Y = aX + b$）
- ・統計ソフトＲ（アール）を使って何かやってみる。

２　量的データと質的データ

（１）量的データ

各データの量を測ることができ、データ同士の足し算、引き算ができる（平均に意味がある）。

（２）質的データ

各データの量を測ることはできず、データ同士の足し算、引き算ができない（平均に意味はない）。

ただし、質的データに変数（ダミー変数）を割り振ることにより、量的データと同様に分析できるようになる（例．男性：0、女性：1）。

データの種類	尺度	特徴	例
量的データ （計量が目的）	間隔尺度	数値の差に意味はあるが、比率に意味はない（ゼロは絶対的なセロではない）。	温度（摂氏） 西暦
	比(率)尺度	数値の差だけでなく、比率にも意味がある（ゼロは絶対的なゼロである）。	温度（華氏） 身長 面積
質的データ （分類が目的）	名義尺度	区別するためだけで、大小関係に意味はない。	男・女 １組・２組
	順序尺度	大小関係に意味はあるが、差に意味はない。	好き・きらい １級・２級・３級

●「割合」は「平均」と同じ？

量的データについては「平均」を計算することができ、質的データについては「割合」を計算することができますが、平均と割合はその本質は同じです。

例えば、5 人の年齢がそれぞれ、10 歳、20 歳、30 歳、40 歳、50 歳だとすると、その平均年齢は（10＋20＋30＋40＋50）／5 人＝30 歳と計算できます。

そして、この 5 人のうち男性が 2 人で女性が 3 人だとすると、女性の割合は 3÷5＝0.6（60%）と計算できます。ここで男性を 0、女性を 1 という数値に置き換えてみると、女性の割合は（0＋0＋1＋1＋1）／5 人＝0.6 ですが、その計算過程を見てみると確かに平均値と同じ計算をしています。

要するに、割合は質的データの平均ということができるということです。

３　平均と分散

（１）平均

$$\bar{X}=\frac{X_1+X_2+\ \cdot\ \cdot\ \cdot+X}{n}=\sum_{i=1}^{n}X_i$$

●平均値と中央値

例えば、5 個の数 1, 2, 3, 4, 10 について考えると、これらの数の平均値は 4、中央値は 3 です。そして、「各数との差の２乗和」を最小にする値は何かというと、これは平均値になります。すなわち、

$(1-4)^2+(2-4)^2+(3-4)^2+(4-4)^2+(10-4)^2=9+4+1+0+36=50$（最小）

ちなみに中央値だと、

$(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(10-3)^2=4+1+0+1+49=55$

となり、確かに 50 より大きな数になります。

一方、「各数との差の絶対値の和」を最小にする値も平均値かと思いきや、この場合は中央値になります。すなわち、

|1－3|+|2－3|+|3－3|+|4－3|+|10－3|=2+1+0+1+7=11（最小）

平均値だと、

|1－4|+|2－4|+|3－4|+|4－4|+|10－4|=3+2+1+0+6=12

となります。どうも私はどちらも正解は平均値のような気がしてしまうのですが、多くの人にとってこれらの結果は自明なことなのでしょうか？

（３）分散

$$\sigma^2=\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(X_{\mathrm{i}}-\bar{\mathrm{X}}\right)^{2}$$

・標準偏差

$$\sigma=\sqrt{\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(X_{\mathrm{i}}-\bar{\mathrm{X}})^2}$$

・不偏分散

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2$$

●不偏分散はなぜｎ－１で割るのか？

これは「各数との差の２乗和（偏差平方和）を最小にする数は平均値だということが分かれば、少なくともｎより小さい数で割るのが妥当だということは理解できると思います。

例えば、母平均が５であることが分かっている正規母集団から「4, 5, 6, 7, 8」という 5 個の標本が得られたとします。

これらのデータの標本平均は 6、偏差平方和は、

$(4-6)^2 + (5-6)^2 + (6-6)^2 + (7-6)^2 + (8-6)^2 = 10$（最小）

となります。そして、これを標本数 5 で割ると標本分散が求まります（この標本の場合は 2)。しかし、母平均は 5 なので、標本平均の代わりに母平均 5 を使ってこの標本偏差平方和を求めると 15 となり、10 より大きくなります（というか、標本平均から少しでもズレた数を使うとその偏差平方和は必ず 10 より大きくなります。)。

もちろん、たまたま「4, 4, 5, 6, 6」のような標本が得られれば標本平均も母平均も 5 になりますが、この場合も標本平均を使った標本の偏差平方和が最小であるという結果は変わりません。ということは、標本平均を使った標本の偏差平方和は、たまたま標本平均が母平均と一致した場合以外は母集団の偏差平方和を過少に見積もってしまうことになるので、不偏分散を求めるためには少なくともｎより小さい数で割ることで調整するのが妥当ということになります。

では、なぜこれがｎ－１なのかという点については、数式展開が必要なので別に機会があればということにしたいと思います。

３　確率分布

（１）正規分布

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

例.　測定誤差、身長、大学入学共通テスト　など

・標準正規分布

$$g(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

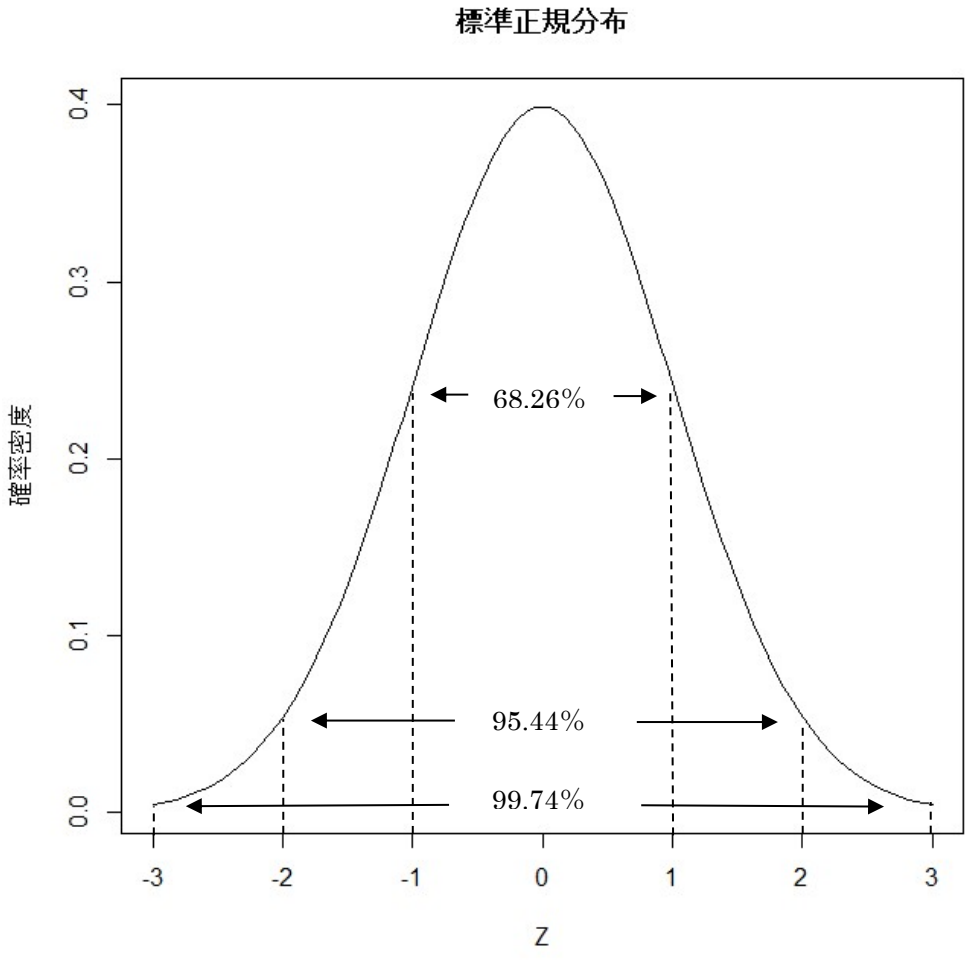
ある観測データの母集団分布を正規分布と仮定すると、
・1 標準偏差の範囲（$\mu-\sigma\sim\mu+\sigma$）に 68.26%のデータが含まれる。
・2 標準偏差の範囲（$\mu-2\sigma\sim\mu+2\sigma$）に 95.44%のデータが含まれる。
・3 標準偏差の範囲（$\mu-3\sigma\sim\mu+3\sigma$）に 99.74%のデータが含まれる。

○正規分布するデータを線形変換$(Y=aX+b)$しても正規分布　→Xが正規分布$N(\mu,\sigma^2)$に従うとき、$aX+b$は$N(a\mu+b,a^2\sigma^2)$に従う。

標準化（基準化）・・・各データの数値から平均値を引き（平均を 0 にする）、標準偏差で割る（標準偏差を 1 にする）。

$$Z=\frac{X-\mu}{\sigma}$$

標準化すると、・Z=－1～＋1 の範囲に 68.26%のデータが含まれる。
・Z=－2～＋2 の範囲に 95.44%のデータが含まれる。
・Z=－3～＋3 の範囲に 99.74%のデータが含まれる。
・Z=－1.96～＋1.96 の範囲に 95.0%のデータが含まれる。



●世の中の数多くの事象はなぜ正規分布するのか？

例えば、表と裏の出る確率がそれぞれ 2 分の 1 ずつのコインを n 回投げて x 回表が出る確率は、二項分布 $_nC_x = \left(\frac{1}{2}\right)^n \left(\frac{1}{2}\right)^{n-x}$ に従い、表の出る回数 x の組合せの数を下記のような図に表すことができます（パスカルの三角形）。

投げる回数									
１回				１	１				
２回			１	２	１				
３回		１	３	３	１				
４回	１	４	６	４	１				
５回	１	５	10	10	５	１			
６回	１	６	15	20	15	６	１		
７回	１	７	21	35	35	21	７	１	
８回	１	８	28	56	70	56	28	８	１

この図は、例えば、

・１回投げた時：表が１回$\left({}_1C_1 = 1 \right)$、表が０回$\left({}_1C_0 = 1 \right)$ →「１，１」

・２回投げた時：表が２回$\left({}_2C_2 = 1 \right)$、表が１回$\left({}_2C_1 = 2 \right)$、表が０回$\left({}_2C_0 = 1 \right)$ →「１，２，１」・・・という感じです。

そして投げる回数 n を増やしていくと次第に正規分布に近づき、n を無限大にして、例えば１標準偏差の範囲で積分すると 0.6826 になるということになります。

ちなみに８回投げたところで、真ん中の 70 とその左右の 56 を足した辺りが１標準偏差内に近いのではないかとあたりをつけて計算してみると、(56＋70＋56)/(1＋8＋28＋56＋70＋56＋28＋8＋1)=182/256≒0,71 となり、確かに 0.6826 に近い値になりました。

○正規母集団からの標本平均の分布

平均μ、分散σ^2の正規分布する母集団から無作為抽出された大きさ n の標本の平均$\overline{X}$は、平均μ、分散σ^2/nの正規分布に従う。

●標本の平均の平均？

正規分布する母集団から 1 個の標本を取り出して元に戻し、また 1 個の標本を取り出して元に戻すという作業を繰り返すと、1 個の標本が数多く得られますが、これらの標本が従う分布は、平均μ、分散σ^2の正規分布です。

また、正規分布する母集団から 10 個の標本を取り出してこれらの平均をとって元に戻し、また 10 個の標本を取り出して平均をとって元に戻すという作業を繰り返すと、10 個の標本平均が数多く得られますが、これらの標本平均が従う分布は、平均μ、分散$\sigma^2/10$の正規分布です。

さらに、正規分布する母集団から 100 個の標本を取り出してこれらの平均をとって元に戻し、また 100 個の標本を取り出して平均をとって元に戻すという作業を繰り返すと、100 個の標本平均が数多く得られますが、これらの標本平均が従う分布は、平均μ、分散$\sigma^2/100$の正規分布です。

「大きさ n の標本の平均$\overline{X}$は、平均μ、分散σ^2/nの正規分布に従う。」ということは、標本の平均の平均がμで、標本の平均の分散がσ^2/nだということを言っています。

○非正規母集団からの標本平均の分布・・・中心極限定理

母集団がどのような分布であっても、母集団から無作為抽出されたｎ個の標本の平均$\bar{X}$は、ｎが大きくなるにつれて、平均μ、分散σ^2/nの正規分布に近づく。

●平均値は正規分布したがっている？

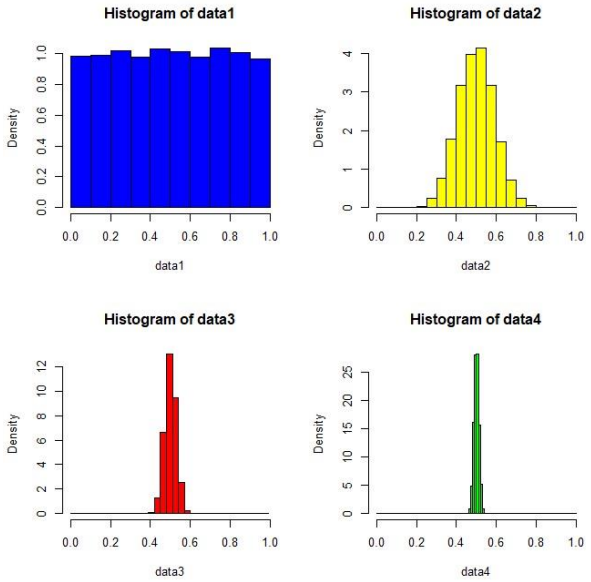
母集団がどのような分布であっても、標本数ｎを大きくすれば、これらのｎ個の標本が正規分布するように誤解してしまいますが、そうではなく、例えば一様分布する母集団から 1 個の標本を取り出して元に戻し、また 1 個の標本を取り出して元に戻すという作業を繰り返すと、1 個の標本が数多く得られますが、これらの標本が従う分布はやはり一様分布です。

しかし、一様分布する母集団から 10 個の標本を取り出してこれらの平均をとって元に戻し、また 10 個の標本を取り出して平均をとって元に戻すという作業を繰り返すと、10 個の標本平均が数多く得られますが、これが平均μ、分散$\sigma^2/10$の正規分布に近づき、さらに取り出す標本数を 100 個にすると平均μ、分散$\sigma^2/100$の正規分布に近づくということを言っています。

○R（アール）で中心極限定理をシミュレーションしてみよう。

一様分布する母集団から、それぞれ 1 個,10 個,100 個,500 個の標本を取り出して、それぞれの標本平均を計算するという作業をそれぞれ 10000 回繰り返す。

```
par(mfrow=c(2,2))
data1<-replicate(10000,mean(runif(1,0,1)))    #1 個の標本抽出を 10000 回繰返し
data2<-replicate(10000,mean(runif(10,0,1)))   #10 個の標本抽出・標本平均の計算を 10000 回繰返し
data3<-replicate(10000,mean(runif(100,0,1)))  #100 個の標本抽出・標本平均の計算を 10000 回繰返し
data4<-replicate(10000,mean(runif(500,0,1))) #500 個の標本抽出・標本平均の計算を 10000 回繰返し
hist(data1,prob=TRUE,breaks=seq(0,1,0.1),xlim=c(0,1),col="blue")
hist(data2,prob=TRUE,breaks=seq(0,1,0.05),xlim=c(0,1),col="yellow")
hist(data3,prob=TRUE,breaks=seq(0,1,0.03),xlim=c(0,1),col="red")
hist(data4,prob=TRUE,breaks=seq(0,1,0.01),xlim=c(0,1),col="green")
```



（２）χ^2分布、ｔ分布

①χ^2分布

確率変数Z_1, Z_2，・・・,Z_nが互いに独立に標準正規分布Ｎ(0, 1)に従うとき、

$\chi^2 = Z_1^2 + Z_2^2 +$ ・・・$+ Z_n^2$の従う分布を自由度ｎのχ^2分布という。

ア　定義より、$N(\mu, \sigma^2)$に従う互いに独立な確率変数をそれぞれ標準化した$\frac{X_1-\mu}{\sigma}, \frac{X_2-\mu}{\sigma}$，・・・,$\frac{X_n-\mu}{\sigma}$の各項の２乗和$\left(\frac{X_1-\mu}{\sigma}\right)^2 + \left(\frac{X_2-\mu}{\sigma}\right)^2 +$ ・・・$+ \left(\frac{X_n-\mu}{\sigma}\right)^2$は、自由度ｎの$\chi^2$分布に従う。

イ　$s^2 = \frac{(X_1-\bar{X})^2+(X_2-\bar{X})^2+\cdots+(X_n-\bar{X})^2}{n-1}$の両辺にそれぞれｎ−1をかけて$\sigma^2$で割ると、

$\frac{(n-1)s^2}{\sigma^2} = \left(\frac{X_1-\bar{X}}{\sigma}\right)^2 + \left(\frac{X_2-\bar{X}}{\sigma}\right)^2 +$ ・・・$+ \left(\frac{X_n-\bar{X}}{\sigma}\right)^2$となるが、これは自由度ｎ−1の$\chi^2$分布に従う。

②ｔ分布

独立な２つの確率変数ＺとＹがあり、Ｚが標準正規分布Ｎ(0, 1)、Ｙが自由度ｎのχ^2分布に従うとき、$t = \frac{Z}{\sqrt{\frac{Y}{n}}}$の従う分布を自由度ｎのｔ分布という。

$Z = \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}$の母分散$\sigma^2$を標本分散$s^2$に置き換えた$t = \frac{\bar{X}-\mu}{\sqrt{\frac{s^2}{n}}}$は、自由度 ｎ−1 のｔ分布に従う。

○統計量ｔのｔ分布の定義式への式変形

$$t = \frac{\bar{X}-\mu}{\sqrt{\frac{s^2}{n}}} = \frac{\frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}}{\frac{\sqrt{\frac{s^2}{n}}}{\sqrt{\frac{\sigma^2}{n}}}} = \frac{Z}{\sqrt{\frac{s^2}{\sigma^2}}} = \frac{Z}{\sqrt{\frac{\frac{(n-1)s^2}{\sigma^2}}{n-1}}} = \frac{Z}{\sqrt{\frac{Y}{n-1}}}$$

●統計量ｔの式変形の意味

統計量Ｚの母分散σ^2の値が分かればＺを利用して平均値の推定・検定（下記４）ができますが、残念ながら通常この値は分かりません。仕方がないので、これを不偏分散s^2に置き換えた統計量ｔで代用します。

このｔを上記のように式変形するとｔ分布の定義式になります。

この式変形は、要するにσ^2をs^2に置き換えた新たな統計量ｔが標準正規分布とどう違う分布をするかを調べるために、ｔの分子を標準正規分布に従うＺに変形してみると、それに伴って分母はχ^2分布に従う分布になります。この標準正規分布をχ^2分布で割った値が従う分布を新たにｔ分布と名付けたということではないでしょうか。

ｔ分布はσ^2（定数）をs^2（確率変数）で代用する分だけ精度が落ちるため、標準正規分布よりやや横に拡がった分布に従いますが、ｎ（標本サイズ）が大きくなるにつれて次第に標準正規分布に近づいていきます。

4　推定と検定

標本のデータから母集団に関する知識を得る方法（母集団の確率分布として正規分布を仮定）

（１）推定

標本のデータから、パラメータ（母数）を推測すること（点推定と区間推定）。

①母平均の区間推定

〔例題 1〕・・・母分散が既知の場合

日本人成人男性 30 人を無作為に抽出して身長を測定したところ、平均値は 170ｃｍであった。この場合の日本人成人男性の平均身長の 95%信頼区間を求めよ。ただし、母分散 は 25 であることが分かっているものとする。

〔解答 1〕

$$-1.96 \leq Z = \frac{\overline{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{170 - \mu}{\sqrt{\frac{25}{30}}} \leq +1.96$$

$$170 - 1.96\sqrt{\frac{25}{30}} \leq \mu \leq 170 + 1.96\sqrt{\frac{25}{30}}$$

$$166.58cm \leq \mu \leq 173.42cm$$

〔例題 2〕・・・母分散が未知の場合（大標本の場合）

日本人成人男性 30 人を無作為に抽出して身長を測定したところ、下記のとおりであった。この場合の日本人成人男性の平均身長の 95%信頼区間を求めよ。ただし、母分散は未知であるものとする。

165 165 165 168 168 172 171 172 171 168 167 180 168 166 173 176 162 169 174 181 173 161 168 172 180 172 169 168 174 167（平均 170.17、不偏分散 24.28）

〔解答 2〕

$$-1.96 \leq Z = \frac{\overline{x} - \mu}{\sqrt{\frac{s^2}{n}}} = \frac{170 - \mu}{\sqrt{\frac{24.28}{30}}} \leq +1.96$$

$$170 - 1.96\sqrt{\frac{24.28}{30}} \leq \mu \leq 170 + 1.96\sqrt{\frac{24.28}{30}}$$

$$168.24cm \leq \mu \leq 171.76cm$$

```
> x<-rnorm(30,170,5)　＃x に平均 170、標準偏差 5 の正規乱数 30 個を格納
> round(x)　＃小数点一位を四捨五入
 [1] 165 165 165 168 168 172 171 172 171 168 167 180 168 166 173 176 162 169 174 181 173 161
168 172 180 172 169 168 174 167
> mean(round(x))　＃平均
[1] 170.1667
> var(round(x))　＃不偏分散
[1] 24.28161
```

〔例題 3〕・・・母分散が未知の場合（小標本の場合）

日本人成人男性 10 人を無作為に抽出して身長を測定したところ、下記のとおりであった。この場合の日本人成人男性の平均身長の 95%信頼区間を求めよ。ただし、母分散は未知であるものとする。

167 180 168 166 173 176 162 169 174 181（平均 168.5、不偏分散 8.28）

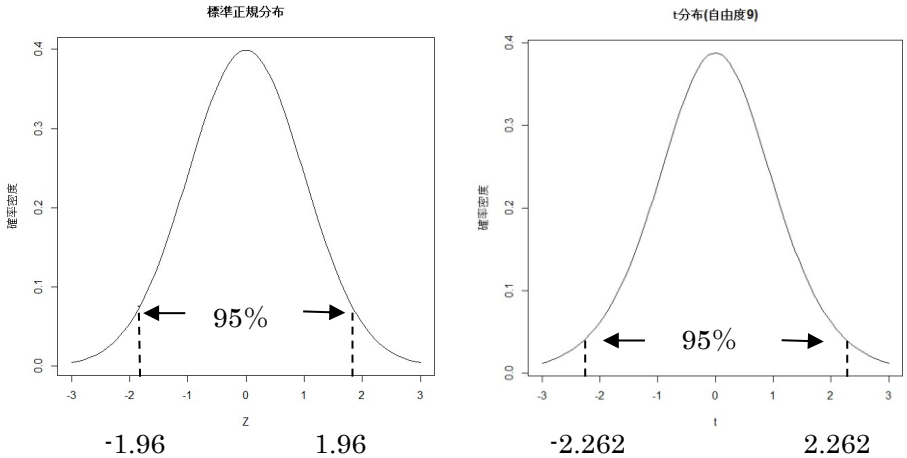
〔解答 3〕

t 分布表より、自由度 n-1=10-1=9 の両側 5%点は、$t_{0.025}(9) = \pm 2.262$である。

$$-2.262 \leq t = \frac{\overline{x} - \mu}{\sqrt{\frac{s^2}{n}}} = \frac{170 - \mu}{\sqrt{\frac{8.28}{10}}} \leq +2.262$$

$$170 - 2.262\sqrt{\frac{8.28}{10}} \leq \mu \leq 170 + 2.262\sqrt{\frac{8.28}{10}}$$

$$167.94cm \leq \mu \leq 172.06cm$$




```
curve(dnorm(x,0,1),-3,3,main="標準正規分布",xlab="Z",ylab="確率密度")　#標準正規分布(Z=-3~3)
curve(dt(x,9),-3,3,main=" t 分布(自由度 9)",xlab="t",ylab="確率密度")　#自由度 9 の t 分布(t=-3~3)
```

（２）検定

母集団の特性についての仮説が正しいかどうかを標本のデータから判断する方法で、採用したい仮説（対立仮説）とこれに相反する仮説（帰無仮説）を設定し、帰無仮説が正しいとすると、観測データが得られる可能性は極めて低いと判断される場合は、帰無仮説を棄却し、対立仮説を採用する。

（手順）

① 仮説（帰無仮説、対立仮説）を設定する。
② 有意水準を設定する。
③ 統計量の実現値が棄却域に入るときは帰無仮説を棄却する。

〔例題〕・・・母分散が未知の場合（小標本の場合）

日本人成人男性 10 人を無作為に抽出して身長を測定したところ、下記のとおりであった。この場合の日本人成人男性の平均身長を 180㎝ と考えてよいか、有意水準 5%で検定せよ。

167 180 168 166 173 176 162 169 174 181（平均 168.5、不偏分散 8.28）

〔解答〕

帰無仮説H_0 : μ =180㎝
対立仮説H_1 : $\mu \neq$180㎝　として両側検定を行う。
t 分布表より、自由度 n-1=10-1=9 の両側 5%点は、$t_{0.025}(9) = \pm 2.262$である。

$$t = \frac{\overline{x} - \mu}{\sqrt{\frac{s^2}{n}}} = \frac{168.5 - 180}{\sqrt{\frac{8.28}{10}}} = -12.64 \leq -2.262$$

上記で求めた t=-12.64 は、-2.262 の外側（棄却域）にある。すなわち、日本人成人男性の平均身長を 180㎝ と仮定すると、平均 168.5㎝ という結果は 5%以下でしか起こりえない。したがって、日本人成人男性の平均身長は 180㎝ という仮説は 5%水準で棄却される。

推薦図書

１　向後千春・冨永敦子「統計学がわかる ファーストブック」（技術評論社）
２　宮川公男「基本統計学［第５版］」（有斐閣）
３　西内啓「統計学が最強の学問である[実践篇]」（ダイヤモンド社）
４　西内啓「統計学が最強の学問である[数学篇]」（ダイヤモンド社）
５　東京大学教養学部統計学教室編「統計学入門」（東京大学出版会）
６　長沼伸一郎「経済数学の直観的方法　確率・統計編」（講談社）