```
library(ggplot2)
library(Hmisc)

##########################################################
#######  （5週目）正規分布とt検定 ######################
##########################################################

#-------
# ○○ 5.1 標本調査
#-------

# 日本人の成人男性全体（母集団）の身長の平均値が知りたい.

# 無作為に10人の身長がわかっているとして,
# そのデータをもとに, 母集団の平均値を推定する.

# [標本１]
x1 = c(153,176,168,167,150,153,170,154,195,191)

# [サンプル数]
n1 = length(x1)
#[1] 10

# [標本１の平均値]
xmean1 = mean(x1)
#[1] 167.7

# [標本１の不偏標準偏差]
## 期待値が母集団の標準偏差と同じとなるように調整した偏差
s1 = sqrt(var(x1))
#[1] 15.97255

## var(x1)は以下の計算と同義
##（分散計算の分母は「n1-1」であることに注意）
sum((x1-xmean1)^2) / (n1-1)

# [標本誤差]
## 標本から推定される母集団平均値分布の標準偏差
se1 = s1 / sqrt(n1)
#[1] 5.050963

# [95%信頼区間]
## 母集団平均が95%の確率で存在する区域（上側境界と下側境界）
myu_max = xmean1 + se1 * qt(0.975, df=n1-1)
#[1] 179.1261
myu_min = xmean1 - se1 * qt(0.975, df=n1-1)
#[1] 156.2739

## t分布の累積確率が97.5%となるt値
## 標準誤差を1としたときの倍率に対応する
```

```
qt(0.975, df=n1-1)
#[1] 2.262157


# X1の分布・不偏標準偏差・標本誤差・95%信頼区間を描画する
library(ggplot2)
gp1 =
  ggplot(NULL,aes(x="X1",y=x1)) +
  geom_boxplot(width=0.2) +
  geom_dotplot(binaxis="y",
binwidth=1,fill="white",colour="black",dotsize=2) +
  scale_x_discrete(name="") +
  scale_y_continuous(name="HEIGHT",limits=c(150,200)) +
  # 標本の不偏標準偏差（s1）の描画
  stat_summary(fun.data="mean_sdl", fun.args = list(mult=1),
               geom="pointrange", size = 2,
               position = position_nudge(0.25),
               colour = "black", shape = 23,fill = "red") +
  # 標本誤差（se1）の描画
  stat_summary(fun.data="mean_se", geom="pointrange", size = 2,
               position = position_nudge(0.5),
               colour = "black", shape = 23,fill = "blue") +
  # 母平均の95%信頼区間の描画
  stat_summary(fun.data="mean_cl_normal", geom="pointrange", size =
2,
               position = position_nudge(0.75),
               colour = "black", shape = 23,fill = "green") +
  theme(text=element_text(size=30))
gp1



# [標本１]に加えて
# 新たに[標本２]を合わせてデータフレーム下し、
# 同様にグラフ化します。
x1 = c(153,176,168,167,150,153,170,154,195,191)
x2 = c(179,176,166,167,170,164,170,154,169,164)

dat = data.frame(height=c(x1,x2),
id_sample=factor(c(rep(1,10),rep(2,10))))

ggplot(dat,aes(x="sample",y=height)) +
  geom_boxplot(width=0.2) +
  geom_dotplot(binaxis="y",
binwidth=1,fill="white",colour="black",dotsize=2) +
  scale_x_discrete(name="") +
  scale_y_continuous(name="HEIGHT",limits=c(150,200)) +
  # 標本の不偏標準偏差（s1）の描画
  stat_summary(fun.data="mean_sdl", fun.args = list(mult=1),
               geom="pointrange", size = 2,
               position = position_nudge(0.25),
               colour = "black", shape = 23,fill = "red") +
```

```
  # 標本誤差（se1）の描画
  stat_summary(fun.data="mean_se", geom="pointrange", size = 2,
               position = position_nudge(0.5),
               colour = "black", shape = 23,fill = "blue") +
  # 母平均の95%信頼区間の描画
  stat_summary(fun.data="mean_cl_normal", geom="pointrange", size =
2,
               position = position_nudge(0.75),
               colour = "black", shape = 23,fill = "green") +
  # 2つのサンプルの統計量を並べて描画
  facet_grid(. ~ id_sample) +
  theme(text=element_text(size=30))

## 標本の分散が少ない標本2の方が、
## 母集団の平均値を精度良く推定できることを確認してください.




#-------
# ○○ 5.2 正規分布とt分布
#-------



#------------------
# 5.2.0 関数描画
#------------------

# まず関数を描画する方法を学びます.


# [関数] curve(xを含む関数式, xmin,xmax)
# [仕様] xを含む関数式のグラフをxminからxmaxの範囲で描画
curve(x^2,-3,3)

# 以下はggplot2を使った方法

library(ggplot2)

# x^2を描画
## fun変数に適用する関数を代入します。
## 以下に示すように2通りの方法がありますが、
## 以下ではgeom_functionを使っていきます。
ggplot() + geom_function(fun=function(x){x^2}, size=1)
ggplot() + stat_function(fun=function(x){x^2}, size=1)

## {}は省略することもできます（というか普通は省略）。
ggplot() + geom_function(fun=function(x) x^2, size=1)
```

```
# x^2とx^3を描画、範囲は-3<x<3
# x^3は赤色、点線
ggplot() +
  geom_function(fun=function(x) x^2, colour="black") +
  geom_function(fun=function(x){x^3-4},
colour="red",linetype="dotdash") +
  scale_x_continuous(limits=c(-3,3))


# x=0、y=0に実線を、適当な交点に点線を引く、
## geom_vline、geom_hlineに、
## エステティック変数：(xy)interceptを指定します。
ggplot() +
  geom_function(fun=function(x) x^2, colour="black") +
  geom_function(fun=function(x){x^3-4}, colour="red") +
  scale_x_continuous(limits=c(-5,5),breaks=seq(-5,5,by=1),name="X")
+
  scale_y_continuous(limits=c(-5,5),breaks=seq(-5,5,by=1),name="Y")
+
  geom_vline(aes(xintercept = c(4^(1/3),2)),linetype="dotdash") +
  geom_hline(aes(yintercept = c(4^(2/3),4)),linetype="dotdash") +
  geom_vline(aes(xintercept = 0),size=1.5) +
  geom_hline(aes(yintercept = 0),size=1.5) +
  theme(text = element_text(size=20))


#------------------
# 5.2.1 正規分布
#------------------

# [関数] dnorm(x,a,b)
# [仕様] 平均a,標準偏差bの正規分布の(x=)xにおける確率密度の値

# 平均170、標準偏差6の正規分布の概形（17歳男性の身長近似）
curve(dnorm(x,170,6),140,200)

# ggplot2で描画
ggplot() + geom_function(fun=function(x){dnorm(x,170,6)}) +
  scale_x_continuous(limits=c(140,200))

# 以下も同様（meanは平均、sdは標準偏差）
ggplot() + geom_function(fun=dnorm, args=list(mean=170,sd=6)) +
  scale_x_continuous(limits=c(140,200))
# argsには関数引数の名前属性を指定していることに注意
args(dnorm)
#function (x, mean = 0, sd = 1, log = FALSE)

# xarray（152,158,164,170,...）のところで縦線を引く
xarray = x=seq(152,188,by=6)
#[1] 152 158 164 170 176 182 188
ggplot() + geom_function(fun=function(x){dnorm(x,170,6)}) +
```

```
  scale_x_continuous(limits=c(140,200),breaks=xarray,name="X") +
  geom_vline(aes(xintercept = xarray),linetype="dotdash") +
  theme(text = element_text(size=30))

options(digits=5) #5桁表示（デフォルトは7桁）

# 確率密度（平均値170cm付近に6.6%、10cm離れると1%を切る）
data.frame(x=xarray,y=dnorm(xarray,170,6))
#    x          y
#1 152 0.00073864
#2 158 0.00899849
#3 164 0.04032845
#4 170 0.06649038
#5 176 0.04032845
#6 182 0.00899849
#7 188 0.00073864


# 164cmから176cmの間に約68%（a-b～a+b）
sum(dnorm(seq(164,176,by=0.1),170,6)*0.1) #[1] 0.68671

# 158cmから182cmの間に約95%（a-2b～a+2b）
sum(dnorm(seq(158,182,by=0.1),170,6)*0.1) #[1] 0.95539

# 152cmから188cmの間に約99%（a-3b～a+3b）
sum(dnorm(seq(152,188,by=0.1),170,6)*0.1) #[1] 0.99737


# 正規分布の性質
# 平均値±標準偏差の範囲に約68%（偏差値40~偏差値60）
# 平均値±2*標準偏差の範囲に約95%（偏差値30～偏差値70）
# 平均値±3*標準偏差の範囲に約99.7%（偏差値20～偏差値80）

# a（平均値）とb（標準偏差）をどのような値に変えても成立します。
a = 70.0; b = 10.0;
sum(dnorm(seq(a-b,a+b,by=0.1),a,b)*0.1) #[1] 0.68511
sum(dnorm(seq(a-2*b,a+2*b,by=0.1),a,b)*0.1) #[1] 0.95594
sum(dnorm(seq(a-3*b,a+3*b,by=0.1),a,b)*0.1) #[1] 0.99734


# これらの確率は、pnorm関数を使うと正確に確認することができます。

# [関数] pnorm(z得点)
# [仕様] ：正規分布における指定したz値
#          （平均値からのズレ/ 標準偏差）
#          の下側確率

ggplot() +
  geom_function(fun=function(x){pnorm(x)}) +
  geom_function(fun=function(x){dnorm(x,0,1)},linetype="dotdash") +
```

```
  scale_x_continuous(limits=c(-5,5),breaks=seq(-5,5,by=1)) +
  theme(text=element_text(size=20))

pnorm(1) - pnorm(-1) #[1] 0.68269（平均値±標準偏差）
pnorm(2) - pnorm(-2) #[1] 0.9549（平均値±2*標準偏差）
pnorm(3) - pnorm(-3) #[1] 0.9973（平均値±3*標準偏差）

# とくに「平均値±2*標準偏差の範囲に約95%」は非常に重要な性質です。
# （ひとまず偏差値30〜70の間に95%の人が含まれる、と考えてください）

# （正確には、2ではなく1.96が係数となります。）
pnorm(1.96) - pnorm(-1.96) #[1] 0.95


# 次に、正規分布の確率分布から、有限個の標本を生成します。

# [関数] rnorm(n,a,b)
# [仕様] 平均a,標準偏差bの正規分布のサンプルをn個生成します。

options(digits=4) #4桁表示（デフォルトは5桁）
a = 170; b = 6 # 平均170・標準偏差7
sum = 10 # 10個のサンプル
rnorm(sum,a,b)
#[1] 170.0 174.6 167.9 162.6 171.7 163.0 170.0 165.2 164.6 166.2

sum = 500000 # 500000個の標本のヒストグラム

h = hist(rnorm(sum,a,b),breaks=seq(140,200,by=0.1))
sum(h$counts[h$breaks>=164 & h$breaks<=176]) / sum #[1] 0.6873（実行
例）
sum(h$counts[h$breaks>=158 & h$breaks<=182]) / sum #[1] 0.9557（実行
例）

#おおよそ近似できていることがわかる。

# 以下はggplotによる描画
ggplot(NULL,aes(x=rnorm(sum,a,b))) +

geom_histogram(breaks=seq(140,200,by=1),colour="black",fill="white")
+
  theme(text=element_text(size=20))



#------------------
# 5.2.2 t分布
#------------------

# t分布は正規分布よりも裾野がわずかに広がります。
```

```
# この裾野の広がりは n が小さいときにより大きくなります。

# 例えばn=10のときのt分布の95%信頼区間は
# 標準誤差の±2.26個分の区間になります.
##（正規分布の場合、±1.96個分）

# [関数] dt(x,df=m)
# [仕様] 自由度m（=n-1）のt分布の、xにおける確率密度を返します。
m = 10-1
ggplot() + geom_function(fun=function(x){dt(x,df=m)}) +
  scale_x_continuous(limits=c(-4,4)) +
  theme(text = element_text(size=20))

# t分布と標準正規分布の形状（赤色）を比べてみます。
ggplot() +
  geom_function(fun=function(x){dt(x,df=3)}) +
  geom_function(fun=function(x){dt(x,df=6)}) +
  geom_function(fun=function(x){dt(x,df=9)}) +
  geom_function(fun=function(x){dt(x,df=12)}) +
  geom_function(fun=function(x){dnorm(x,0,1)},colour="red") +
  scale_x_continuous(limits=c(-4,4)) +
  theme(text = element_text(size=20))

## nが小さいとピークがつぶれ裾野が広がることがわかります.
## 逆にnが大きくなると, 標準正規分布（赤色）に収束していきます.

# x=-∞からの累積確率がpの確率点を求めるには[qt]を使います。
# [関数] qt(p,df=m)
# [仕様] 自由度m（=n-1）のt分布における、累積確率がpの時のt値を返します。

# 95%信頼区間の場合、左側の累積確率は0.025となります。
# 実際、先ほどの理論値と一致しています。
qt(0.025,df=9) #[1] -2.262
qt(0.975,df=9) #[1] 22.262

# n=4、100、10000のときの95%信頼区間の境界

n = 4
qt(0.025,df=n-1);qt(0.975,df=n-1)
#[1] -3.182446
#[1] 3.182446

n = 100
qt(0.025,df=n-1);qt(0.975,df=n-1)
#[1] -1.984217
#[1] 1.984217

n = 100000
qt(0.025,df=n-1);qt(0.975,df=n-1)
#[1] -1.959988
```

```
#[1] 1.959988

##正規分布の95%ラインとほぼ同じ
qnorm(0.025);qnorm(0.975)
#[1] -1.959964
#[1] 1.959964

qt(0.95,df=9)
#[1] 1.833113

# 以上をグラフ化します（点線は標準正規分布）。

n = 4
qt(0.975,df=n-1)
#[1] 2.262157
ggplot() + geom_function(fun=function(x){dt(x,df=n-1)},size=1.5) +
  geom_function(fun=function(x)
{dnorm(x,0,1)},linetype="dotdash",size=0.5) +

geom_vline(aes(xintercept=c(qt(0.025,df=n-1),qt(0.975,df=n-1))),colo
ur="red") +
  scale_x_continuous(limits=c(-4,4),name=NULL) +
  theme(text = element_text(size=30))

n = 10
qt(0.975,df=n-1)
#[1] 2.262157
ggplot() + geom_function(fun=function(x){dt(x,df=n-1)},size=1.5) +
  geom_function(fun=function(x)
{dnorm(x,0,1)},linetype="dotdash",size=0.5) +

geom_vline(aes(xintercept=c(qt(0.025,df=n-1),qt(0.975,df=n-1))),colo
ur="red") +
  scale_x_continuous(limits=c(-4,4),name=NULL) +
  theme(text = element_text(size=30))

n = 100
qt(0.975,df=n-1)
#[1] 1.984217
ggplot() + geom_function(fun=function(x){dt(x,df=n-1)},size=1.5) +
  geom_function(fun=function(x)
{dnorm(x,0,1)},linetype="dotdash",size=0.5) +

geom_vline(aes(xintercept=c(qt(0.025,df=n-1),qt(0.975,df=n-1))),colo
ur="red") +
  scale_x_continuous(limits=c(-4,4),name=NULL) +
  theme(text = element_text(size=30))




# 逆に検定量のt値がわかっている場合、
# 下側確率は以下で求めることができます。
```

```
# [関数] pt(t0, df=m, lower.tail=TRUE/FALSE)
# [仕様] 自由度m（=n-1）のt分布における、検定量t0に対する
#         下側確率P(t<t0)：lower.tail=T（デフォルト）または
#         上側確率P(t>t0)：lower.tail=F

pt(-2.262,df=9) #[1] 0.02501
pt(2.262,df=9)  #[1] 0.975
pt(2.262,df=9,lower.tail=F) #[1] 0.02501 （上側確率）




#-------
# ○○ 5.3 中心極限定理
#-------

# 母集団からサンプルnで標本を採取することを
# 何回も繰り返して、各標本の平均値の分布をとると、
# それらの分布の平均値と標準偏差がどのような値をとるかを
# 確認しましょう.

# 以下では、無限に1〜6（サイコロ）が一様
# に存在している母集団を考えます。
# ここでは


# サイコロベクトルの定義
dice=1:6

# [母集団の平均値]
# 母集団である（無限）サイコロの平均値
mean(dice) #[1] 3.5

# [母集団の標準偏差]
## 不偏ではないので、n-1ではなくnで割り直します.
sqrt(var(dice) * 5/6)
#[1] 1.708


# 母集団からサンプル数nで標本を抜き出し、
# それらの平均値を計算する関数
getMeanDice = function(n){
  sn = sample(dice,n,replace=T)
  mean(sn)
}

# 引数をベクトルに拡張します
getMeanDiceVector = function(nv){
```

```
  result = vector("double",length(nv))
  for(i in 1:length(nv)){
    sn = sample(dice,nv[i],replace=T)
    result[i] = mean(sn)
  }
  result
}

# nの数を変えて実行
nvector = c(2,4,6,8,10,100,1000,10000)
getMeanDiceVector(nvector)
#[1] 3.000 3.250 2.500 3.625 3.300 3.550 3.451 3.485

#各nについてそれぞれ500回の平均値を集める
narray = rep(nvector,times=500)
marray = getMeanDiceVector(rep(narray,times=500))


# データフレームとする
# N：サイコロを振る回数、M：500試行の平均値
dat = data.frame(N = narray,M = marray)

# 「サイコロをn回振って出た目の平均値」の分布（500試行）！！
# ピークは移動せずも、幅が徐々に先鋭となることに注意！！
ggplot(dat,aes(x=M)) +
  geom_histogram(binwidth=0.05) +
  facet_grid(. ~ N) +
  theme(text=element_text(size=20))


# 標本の平均値の分布の標準偏差を可視化
## サンプルサイズが大きいほど、散らばりが小さくなる
dat = data.frame(M = marray, N = factor(narray))
ggplot(dat,aes(x=N,y=M)) +
  geom_boxplot(width=0.5) +
  stat_summary(fun.data="mean_sdl", fun.args = list(mult=1),
               geom="pointrange", size = 1,
               position = position_nudge(0.5),
             colour = "black", shape = 23,fill = "red") +
  theme(text=element_text(size=30))

## 標本平均分布の標準偏差は「標準誤差」と呼ばれ,
## 「母集団の標準偏差 / √サンプルサイズ」として計算される.
## これを中心極限定理と呼ぶ.

# [中心極限定理]
# サンプルサイズ n が大きいとき
# 標本平均分布の標準偏差は以下で近似できる
# 「母集団の標準偏差」 / sqrt(n)
```

```
# 要するに、サンプルサイズが n 倍となると
# グラフの幅は（1/√n）だけ縮小する。

# 母集団の標準偏差は通常は未知
# そのため、標本の分布から√不偏分散として計算される.
# 不偏分散は「平均からの差の2乗和」/ 「要素数−1」
# 不偏分散はvar(ベクトル)で一発で計算できます。

# [関数] var(標本ベクトル)
# [仕様] 不偏分散（「平均からの差の2乗和」/ 「要素数−1」）を返す

# 母集団（無限にサイコロがある状態）の標準偏差は？
## 無限を10000で代替
sqrt(var(rep(1:6,10000)))
# [1] 1.708

# サンプルサイズが4倍になると、標準偏差が約1/√4（半分）となることが確認できる
## 注意）結果は環境によって異なります。
sqrt(var(dat$M[dat$N==2]));
#[1] 1.228
sqrt(var(dat$M[dat$N==4]));
#[1] 0.8545
sqrt(var(dat$M[dat$N==6]));
#[1] 0.6992
sqrt(var(dat$M[dat$N==8]));
#[1] 0.5802


# サンプルサイズが100倍になると、標準偏差が1/√100（1/10）となることが確認できる
sqrt(var(dat$M[dat$N==10]));
#[1] 0.5257
sqrt(var(dat$M[dat$N==100]));
#[1] 0.1682
sqrt(var(dat$M[dat$N==1000]));
#[1] 0.05232
sqrt(var(dat$M[dat$N==10000]));
#[1] 0.01676


# このサンプルサイズnを大きくしていくと
# 標本平均分布は「正規分布」に収束していく！！
##（ここではhist関数でグラフを表示させます）

hist(dat$M[dat$N==100],breaks=seq(1,6,by=0.05),xlim = c(3.0,4.0))
hist(dat$M[dat$N==1000],breaks=seq(1,6,by=0.01),xlim = c(3.3,3.7))
hist(dat$M[dat$N==10000],breaks=seq(1,6,by=0.001),xlim = c(3.4,3.6))

################################################################
# 「正規分布」とは、任意の分布の標本平均分布と関係しています。
# もともとの分布が正規分布である必要はありません。
```

```
# あらゆる分布の（標本平均としての）メタ分布が「正規分布」なのです。
# この意味をよくよく噛み締めてください。
#####################################################################




#---------
# ○○○○ 5.4 標本誤差と信頼区間
#---------


# [標本誤差分布]----------
# 標本誤差とは、母集団の平均値と標本の平均値の誤差を指し、
# 標本誤差分布とは、この標本誤差の分布を指します。


# 平均170、標準偏差6の正規分布から、サイズ100の標本を抽出
a = 170; b = 6; n = 100
h.100 = replicate(10000,mean(rnorm(n,a,b))) #10000の標本平均を集めます

# 標本平均の分布（n=100）
ggplot(NULL,aes(x=h.100)) +

geom_histogram(breaks=seq(165,175,by=0.1),colour="black",fill="white
")

# 標本誤差の分布
## 標本誤差は標本平均から、母集団の平均を引いたものです。
ggplot(NULL,aes(x=h.100-a)) +

geom_histogram(breaks=seq(-5,5,by=0.1),colour="black",fill="white")

# 標本誤差分布の平均はゼロとなり、
# グラフの幅は母平均の標準偏差より狭くなっていることがわかります。

# 標本誤差分布は、標本平均分布と同様に、
# 標本数を増やしていくと、正規分布に近づきます。

# また、標本誤差分布の標準偏差は、
# 標本平均分布の標準偏差と同様に以下で計算されます。
# b / √n ｜（母集団の標準偏差）/ √標本サイズ

b / sqrt(100) #[1] 0.6

# 母集団が未知の場合、標本誤差分布の標準偏差は、
# bを（特定の）標本の不偏標準偏差sに変えて計算します。
```

```
s = sqrt(var(rnorm(100,a,b))); s/sqrt(100) #[1] 0.5215
s = sqrt(var(rnorm(100,a,b))); s/sqrt(100) #[1] 0.6168
s = sqrt(var(rnorm(100,a,b))); s/sqrt(100) #[1] 0.6299

# ばらつきはありますが、おおよそいい推定ができています。



#-----------
# ○○○○○ 5.5 t検定
#-----------

#------------------
# 5.5.1 1標本のt検定
#------------------

# アメリカ人の成人男性の平均身長は175cm（2001年調べ）です。
# 日本人の成人男子の身長が、
# アメリカ人の平均身長と比較して違いがあるかを検定します。

# [帰無仮説] 日本人の成人男性の母集団平均は175cmである。
# [対立仮説１] 日本人の成人男性の母集団平均は175cmではない。（両側検定）
# [対立仮説２] 日本人の成人男性の母集団平均は175cmより低い。（下側片側検定）

# これを以下の2つのサンプルで考えましょう。


n = 10 # サンプルサイズ
x1 = c(179,176,166,167,170,164,170,154,169,164) # 標本1
x2 = c(153,176,168,167,150,153,170,154,195,191) # 標本2

# データフレーム化します
dat = data.frame(GROUP=c(rep(1,10),rep(2,10)),HEIGHT=c(x1,x2))

#箱ひげ図のplot（平均値とアメリカの平均値も描画）
ggplot(dat,aes(x=factor(GROUP),y=HEIGHT)) +
  geom_boxplot() +
  stat_summary(fun="mean", geom="point", shape = 21, size = 8., fill
= "red") +
  geom_dotplot(binaxis="y", binwidth=1) +
  geom_hline(aes(yintercept=175),colour="purple",linetype="dotdash")
+
  theme(text=element_text(size=20))


# 標本平均（xmean1 > xmean2）
xmean1 = mean(dat$HEIGHT[dat$GROUP==1]) #167.9
xmean2 = mean(dat$HEIGHT[dat$GROUP==2]) #167.7

# 標本の不偏標準偏差（s1 < s2）
```

```
s1 = sqrt(var(x1)) #6.887
s2 = sqrt(var(x2)) #15.97

# サンプル1のt値は
t1 = (xmean1 - 175) / (s1/sqrt(n)) #[1] -3.26
# サンプル2のt値は
t2 = (xmean2 - 175) / (s2/sqrt(n)) #[1] -1.445


# [サンプル1のt検定]----------

# t<t1となる確率p1は（p値）？
p1 = pt(t1,df=9) #[1] 0.004919 （下側約0.5%の領域）

# 母平均が175cmのときに、
# サンプル1以上に下側に偏る確率は0.492%（片側検定）
# サンプル1以上のズレが生じる確率は0.984%（両側検定）

# これらはいずれについても5%を下回る。よって、

# 帰無仮説
# 「（日本人成人男性の）母集団平均は175cm」である
# は棄却され、
# 対立仮説１「（日本人成人男性の）母集団平均は175cmではない」または
# 対立仮説２「（日本人成人男性の）母集団平均は175cmより低い」が
# 採用される。

# 統計的には、日本人の成人男性の身長は、
# アメリカ人よりも有意に低いと結論できる。



# [サンプル2のt検定]----------

# t<t2となる確率p2（p値）は？
p2 = pt(t2,df=9) #[1] 0.091

# 母平均が175cmのときに、

# サンプル1ほどに下側に偏る確率は9.1%（片側検定）
# サンプル1ほどにズレが生じる確率は18.2%（両側検定）

# いずれについても有意水準の5%を上回る。よって
# 帰無仮説
# 「（日本人成人男性の）母集団平均は175cm」は支持される。

# 統計的には、日本人の成人男性とアメリカ人の成人男性との間に
```

```
# 有意な差は存在しない。

# 自由度9のt分布上で、2つのサンプルのt値がどこにあるかを確認します。

#まずt分布を描画したのちに
#t=t1とt=t2に線を引きます。｜abline(縦線v=x、横線h=y)
ggplot() + geom_function(fun=function(x){dt(x,df=9)}) +
  geom_vline(aes(xintercept=c(t1,t2)),linetype="dotdash") +
  scale_x_continuous(limits=c(-4,4))

############################################################
#! 標本2と標本1では平均値が標本1の方が高いにもかかわらず
#! 標本1でのみ有意な差が検出されていることに注意。
#! 標本2はデータの散らばりが大きいため、
#! 標準偏差で割って標準得点化されるt値では、
#! 元々の平均差が低く査定される。
############################################################



# [Rの関数を用いる]----------

# t.test関数を用いれば、
# 以上の計算を自動的に行ってくれます。

# [関数] t.test(標本ベクトル, mu=比較対象の母平均, alternative=片側or両側)
# [仕様] 標本ベクトルと比較対象の母平均とを比較する1標本t検定
#       alternative="two.sided"（両側検定）
#       alternative="less"（片側検定：下側確率）
#       alternative="greater"（片側検定：上側確率）



#標本１
x1 = c(153,176,168,167,150,
       153,170,154,195,191)

# 標本２
x2 = c(179,176,166,167,170,
       164,170,154,169,164)



# サンプル1のt検定（両側確率）
t.test(x1,mu=175,alternative="two.sided")
#data:  x1
#t = -1.4453, df = 9, p-value = 0.1823
#alternative hypothesis: true mean is not equal to 175
```

```
#95 percent confidence interval:
#  156.2739 179.1261
#sample estimates:
#  mean of x
#167.7


# サンプル1のt検定（下側確率）
t.test(x1,mu=175,alternative="less")
#data:  x1
#t = -1.4453, df = 9, p-value = 0.09114
#alternative hypothesis: true mean is less than 175
#95 percent confidence interval:
#  -Inf 176.959
#sample estimates:
#  mean of x
#167.7




# サンプル2のt検定（下側確率）
t.test(x2,mu=175,alternative="less")
#data:  x2
#t = -3.26, df = 9, p-value = 0.004919
#alternative hypothesis: true mean is less than 175
#95 percent confidence interval:
#  -Inf 171.8924
#sample estimates:
#  mean of x
#167.9


# サンプル2のt検定（両側確率）
t.test(x2,mu=175,alternative="two.sided")
#data:  x2
#t = -3.26, df = 9, p-value = 0.009839
#alternative hypothesis: true mean is not equal to 175
#95 percent confidence interval:
#  162.9732 172.8268
#sample estimates:
#  mean of x
#167.9




#------------
# ○○○○○○○ 5.7 2標本のt検定（対応あり）
#------------

# 同じ被験者のサプリメント
```

```
# 接種前（pre）と接種後（post）の体重の変化

sbj = LETTERS[c(1:8)] #LETTERSは大文字のアルファベットのベクトル
pre = c(95,80,80,85,75,75,80,85)
post = c(90,75,75,75,80,65,75,80)

dat = data.frame(SBJ=c(sbj,sbj),
                 ORDER=c(rep("PRE",8),rep("POST",8)),
                 WEIGHT=c(pre,post))

#箱ひげ図
ggplot(dat,aes(x=ORDER,y=WEIGHT)) +
  geom_boxplot() +
  stat_summary(fun="mean", geom="point", shape = 21, size = 8., fill
= "red")

# サプリメント接種前と接種後で体重に変化があるか（両側検定）
# あるいは体重は減少したか（片側検定：下側）

# 実は、対応のあるt検定は、
# 少しの操作で1標本t検定に変換できる。
# なぜなら、

# 帰無仮説「標本Aと標本Bに差はない」は
# 帰無仮説「標本（A−B）の母集団平均は0と等しい」

# と同義であるためである。

# 同じ被験者のサプリメント
# 接種前（pre）と接種後（post）の体重
pre = c(95,80,80,85,75,75,80,85)
post = c(90,75,75,75,80,65,75,80)
# 接種前（pre）と接種後（post）の体重の変化
change = pre-post
#[1]  5  5  5 10 -5 10  5  5

#片側検定
t.test(change,dm=0,alternative="less")
#data:  change
#t = -3.1, df = 7, p-value = 0.009
#alternative hypothesis: true mean is less than 0
#95 percent confidence interval:
#  -Inf -1.899
#sample estimates:
#mean of x
#       -5

# change（接種後−接種前）の母平均が0であると仮定したときに
# changeが、標本以上の減少を示す確率は0.9%
# 母平均を0とする仮説は棄却され
```

```
# 対立仮説（片側検定）である「changeは0より小さい」が採用される。


# ちなみに、t.testに
# 比較対象の2つのベクトルを引数として入れることもできます。
# この場合、paired=TRUEを指定する必要があります。

t.test(post,pre,paired=TRUE,alternative="less")
#data:  data$post and data$pre
#t = -3.1, df = 7, p-value = 0.009
#alternative hypothesis: true mean is less than 0
#95 percent confidence interval:
#  -Inf -1.899
#sample estimates:
#mean of x
#       -5
```