

統計学 2020 Lecture 11: R を用いた演習 (2)

正規分布におけるパラメータの点推定・区間推定・仮説検定

北門 利英（東京海洋大学海洋生物資源学科）

2020 年 7 月 22 日

目次

1	アジ科の魚の形態に関するデータ	2
1.1	データの説明と図示	2
1.2	データの読み込み	2
1.3	データの図示	3
2	パラメータの推定	7
3	1 変数の正規分布に対する区間推定	8
3.1	区間推定の復習	8
3.2	R を用いた μ の信頼区間の計算 (分散が未知の場合)	8
4	1 変数の正規分布に対する仮説検定	9
4.1	仮説検定の復習	9
4.2	R を用いた μ に対する仮説検定の実施 (分散未知)	10
5	正規分布における 2 標本検定 (興味があれば取り組んでみて下さい)	11
6	提出課題	14

Attention:

- 今回も R による演習を行います。
- 7 月 22 日もリモートで質問を受け付けます (13:00-13:45 とします)。

1 アジ科の魚の形態に関するデータ

1.1 データの説明と図示

\begin{itembox}[l]{例題} 表1は、アジ科の魚であるマアジ、モロ、カイワリそしてマルアジの形態測定データを示している。ここで、TL は全長 (total length), FL は尾叉長 (fork length), SL は標準体長 (standard length), BD は体高 (body depth), HL は頭長 (head length), そして ED は眼径 (eye diameter) の測定値である。いま、体高/標準体長を魚のプロポーシオンとして定義するとき、このプロポーシオンに関する統計的推測を行いたい。 \end{itembox}

この例では、アジ科の4種の個体を市場から仕入れ形態測定を行っている。したがって、例えば海に生息するマアジ個体群からのランダムサンプリングかと言われると必ずしもそうではないため、体長を種間で比較するようなことはできない。しかしながら、ここでは体高/標準体長を測定値として扱うため、産卵から一定期間経過してから体の形が相似形を保って成長すると考えられると仮定し、このデータを母集団からのプロポーシオンのランダムサンプリングと考えてよいこととする。

表1: アジ科の魚の形態測定データ

Species	TL (mm)	FL (mm)	SL (mm)	HL (mm)	BD (mm)	ED (mm)
Maaaji	297	259	235	65	59	15.4
Maaaji	285	263	255	65	61	15.2
						
Maruaji	182	148	162	38	34	7.7

1.2 データの読み込み

```
Data <- read.csv("horsemackerel.csv", header=T)
Data <- cbind(Data, Prop=Data$BD/Data$SL); attach(Data)
head(Data)
```

```
Species TL FL SL HL BD ED Prop
1 Maaaji 297 259 235 65 59 15.4 0.2510638
2 Maaaji 285 263 255 65 61 15.2 0.2392157
3 Maaaji 285 255 263 65 63 15.1 0.2395437
4 Maaaji 319 275 269 71 63 16.6 0.2342007
5 Maaaji 297 259 235 65 59 15.1 0.2510638
6 Maaaji 194 172 160 45 42 13.6 0.2625000
```

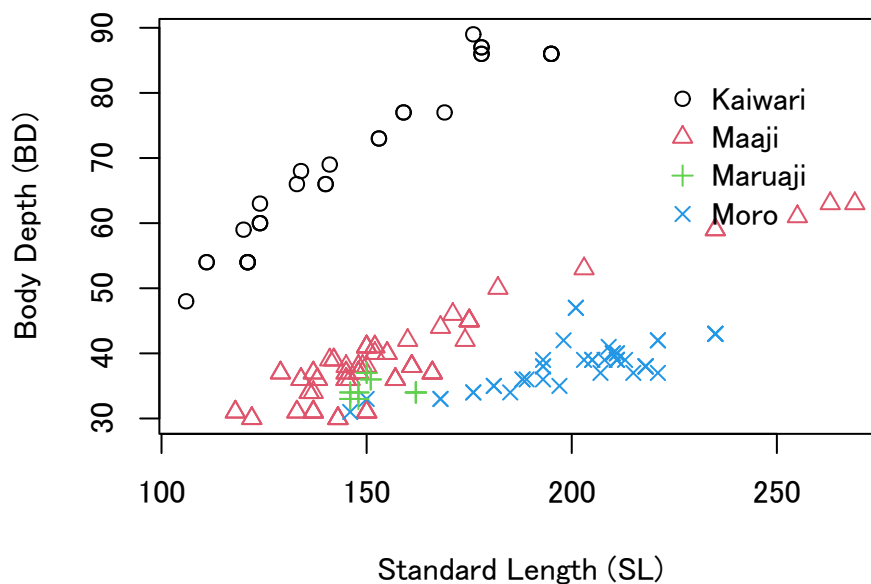
```
tail(Data)
```

	Species	TL	FL	SL	HL	BD	ED	Prop
170	Maruaji	182	162	148	38	33	7.7	0.2229730
171	Maruaji	182	148	162	38	34	7.7	0.2098765
172	Maruaji	182	148	162	38	34	7.7	0.2098765
173	Maruaji	182	148	162	38	34	7.7	0.2098765
174	Maruaji	179	165	150	40	37	10.7	0.2466667
175	Maruaji	182	148	162	38	34	7.7	0.2098765

1.3 データの図示

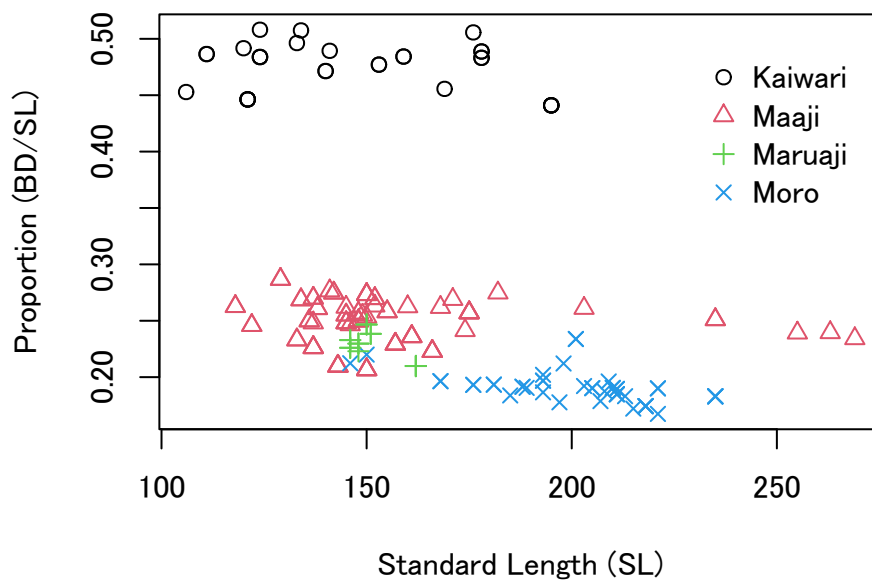
1.3.1 標準体長と体高の関係 (2 次元散布図)

```
SPf <- factor(Species)
SPn <- as.numeric(SPf)
plot(SL, BD, pch=SPn, col=SPn,
      xlab="Standard Length (SL)", ylab="Body Depth (BD)")
legend(220, 85, legend=levels(SPf), pch=1:4, col=1:4, bty="n")
```



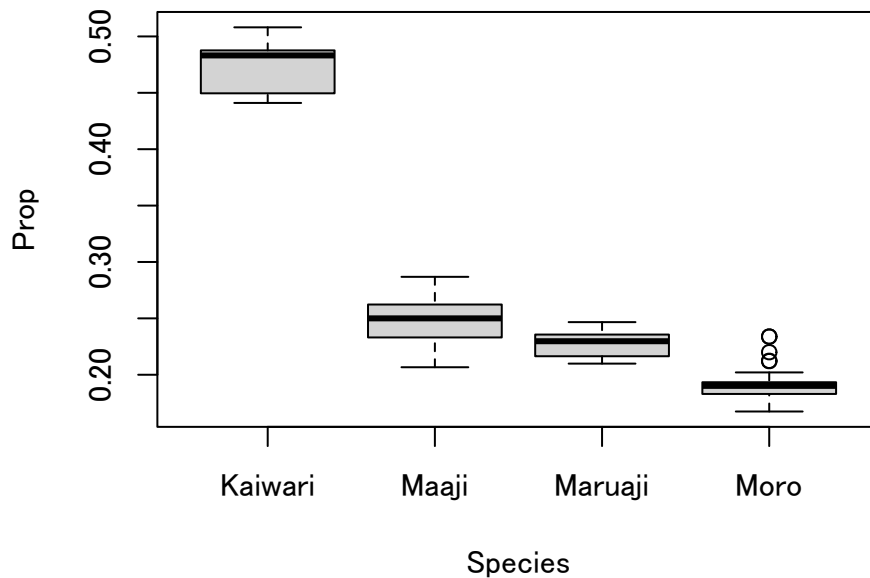
1.3.2 標準体長とプロポーションの関係 (2次元散布図)

```
plot(SL, Prop, pch=SPn, col=SPn,  
      xlab="Standard Length (SL)", ylab="Proportion (BD/SL)")  
legend(230, 0.5, legend=levels(SPf), pch=1:4, col=1:4, bty="n")
```



1.3.3 プロポーションのボックスプロット (箱ひげ図)

```
boxplot(Prop~Species)
```



1.3.4 プロポーションのヒストグラム

種毎にデータを分割

```
y <- split(Prop, Species)
```

```
y
```

```
$Kaiwari
```

```
[1] 0.4893617 0.5074627 0.5056818 0.5080645 0.4916667 0.4556213 0.4528302
[8] 0.4962406 0.4714286 0.4771242 0.4838710 0.4887640 0.4842767 0.4838710
[15] 0.4842767 0.4887640 0.4714286 0.4771242 0.4838710 0.4864865 0.4831461
[22] 0.4410256 0.4462810 0.4410256 0.4462810 0.4410256 0.4462810 0.4831461
[29] 0.4864865 0.4410256 0.4462810
```

```
$Maaji
```

```
[1] 0.2510638 0.2392157 0.2395437 0.2342007 0.2510638 0.2625000 0.2413793
[8] 0.2747253 0.2690058 0.2619048 0.2631579 0.2610837 0.2481752 0.2550336
[15] 0.2533333 0.2482759 0.2465753 0.2482759 0.2500000 0.2567568 0.2700730
[22] 0.2551724 0.2580645 0.2482759 0.2500000 0.2567568 0.2700730 0.2551724
[29] 0.2580645 0.2620690 0.2627119 0.2459016 0.2608696 0.2330827 0.2746479
[36] 0.2868217 0.2686567 0.2697368 0.2262774 0.2097902 0.2292994 0.2066667
```

```
[43] 0.2500000 0.2733333 0.2571429 0.2228916 0.2360248 0.2262774 0.2097902
[50] 0.2292994 0.2066667 0.2500000 0.2733333 0.2571429 0.2228916 0.2360248
[57] 0.2262774 0.2097902 0.2292994 0.2066667 0.2500000 0.2733333 0.2571429
[64] 0.2228916 0.2360248 0.2765957 0.2627119 0.2459016 0.2608696 0.2330827
[71] 0.2746479 0.2868217 0.2686567 0.2697368 0.2262774 0.2097902 0.2292994
[78] 0.2066667 0.2500000 0.2733333 0.2571429 0.2228916 0.2360248
```

```
$Maruaji
```

```
[1] 0.2260274 0.2328767 0.2297297 0.2384106 0.2384106 0.2297297 0.2260274
[8] 0.2328767 0.2466667 0.2229730 0.2098765 0.2098765 0.2098765 0.2466667
[15] 0.2098765
```

```
$Moro
```

```
[1] 0.1902439 0.1830986 0.1720930 0.1674208 0.1904762 0.1865285 0.1776650
[8] 0.2121212 0.1848341 0.1964286 0.1895735 0.1904762 0.1933702 0.1931818
[15] 0.1914894 0.2338308 0.1848341 0.1895735 0.1904762 0.1933702 0.1931818
[22] 0.1914894 0.2338308 0.1964286 0.1848341 0.1900452 0.1968912 0.1829787
[29] 0.1743119 0.1829787 0.1743119 0.1829787 0.1743119 0.1900452 0.1968912
[36] 0.1829787 0.1743119 0.1875000 0.2020725 0.1961722 0.1787440 0.1902439
[43] 0.1921182 0.1837838 0.2200000 0.2123288
```

```
# ヒストグラムの作成
```

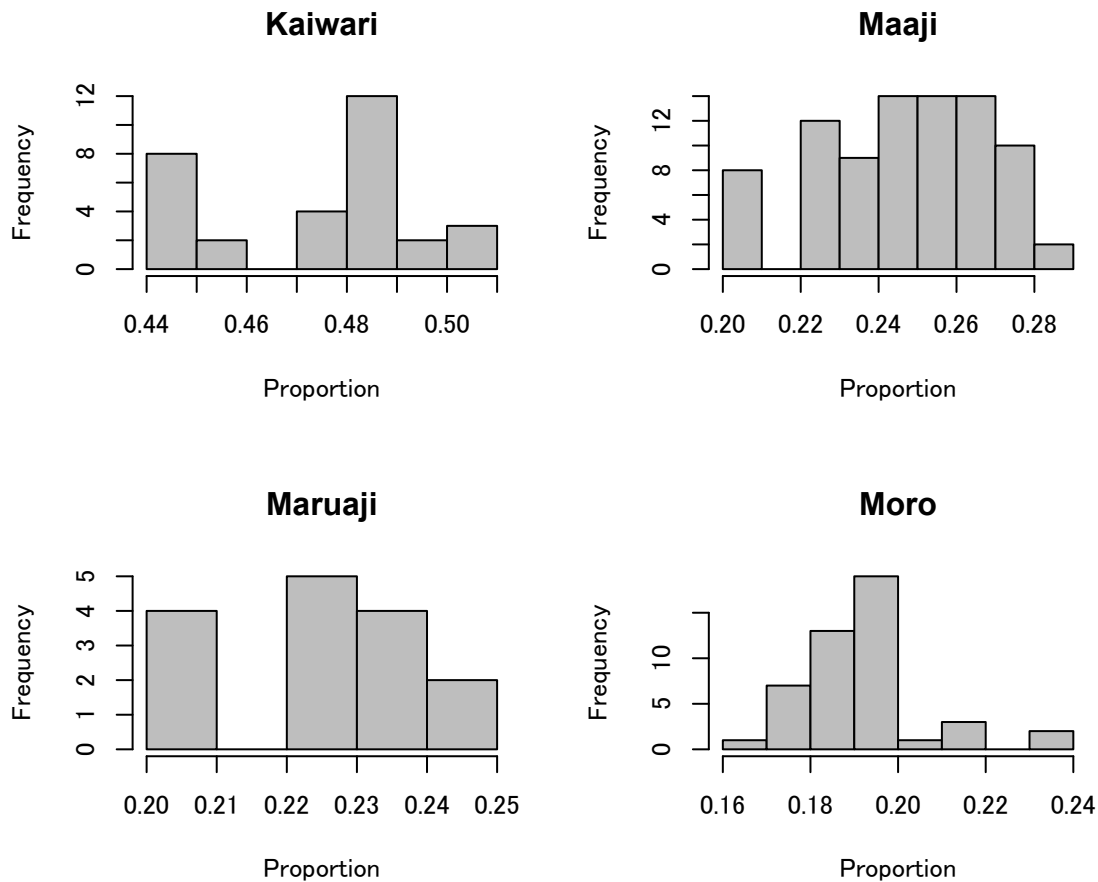
```
par(mfrow=c(2,2)) # 画面を 2*2 に分割
```

```
hist(y$Kaiwari, main="Kaiwari", col="gray", xlab="Proportion")
```

```
hist(y$Maaji, main="Maaji", col="gray", xlab="Proportion")
```

```
hist(y$Maruaji, main="Maruaji", col="gray", xlab="Proportion")
```

```
hist(y$Moro, main="Moro", col="gray", xlab="Proportion")
```



2 パラメータの推定

Lecture 7 の資料でまず復習してください。

```
# Estimation of mu
Ns <- sapply(y, length) # 種ごとのデータの数
Mean <- sapply(y, mean) # 種ごとのプロポーションの標本平均の値
Var <- sapply(y, var)   # 種ごとのプロポーションの標本 (不偏) 分散の値
SD <- sapply(y, sd)     # 種ごとのプロポーションの標準偏差の値
SE.mu <- SD/sqrt(Ns)    # 種ごとの mu の標準誤差 (推定量の標準偏差)
data.frame(
  Nsample=Ns,
  Mean=round(Mean,3),
  Variance=round(Var,4),
  SD=round(SD,3),
  SE.mu=round(SE.mu,4)
```

)

	Nsample	Mean	Variance	SD	SE.mu
Kaiwari	31	0.474	5e-04	0.022	0.0039
Maaji	83	0.248	4e-04	0.020	0.0022
Maruaji	15	0.227	2e-04	0.013	0.0033
Moro	46	0.191	2e-04	0.014	0.0020

3 1 変数の正規分布に対する区間推定

3.1 区間推定の復習

Lecture 8 の資料でまず復習してください.

標準偏差 σ が既知と考えられる場合には,

$$\bar{Y} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{Y} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad (1)$$

で信頼区間を求めることができる.

ただし, 通常は標準偏差 σ が未知であり, $\hat{\sigma}^2$ を代わりに用いるが, $\hat{\sigma}^2$ はあくまで推定値であり不確実性が伴うから式 (??) が成り立たず, 自由度 $(n-1)$ の t 分布に従う.

$$T = \frac{\bar{Y} - \mu}{\hat{\sigma}/\sqrt{n}} \sim t(n-1)$$

$$\bar{Y} - t_{\alpha/2}(n-1) \frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{Y} + t_{\alpha/2}(n-1) \frac{\hat{\sigma}}{\sqrt{n}}$$

が水準 $(1-\alpha)100\%$ の信頼区間となる. ただし, $t_{\alpha/2}(n-1)$ は自由度 $(n-1)$ の上側 $\alpha/2 \times 100\%$ 点である.

3.2 R を用いた μ の信頼区間の計算 (分散が未知の場合)

実際にアジ科の魚のプロポーションの信頼区間を求めてみると以下ようになる.

3.2.1 自分で式をつくって計算する場合

```
# 4 種同時に求めています
alpha <- 0.05
LB <- Mean - SD/sqrt(Ns)*qt(1-alpha/2, Ns-1)
UB <- Mean + SD/sqrt(Ns)*qt(1-alpha/2, Ns-1)
data.frame(
  Mean=round(Mean,3),
```

```
LB=round(LB,4),
UB=round(UB,4)
)
```

	Mean	LB	UB
Kaiwari	0.474	0.4660	0.4818
Maaji	0.248	0.2433	0.2523
Maruaji	0.227	0.2202	0.2344
Moro	0.191	0.1867	0.1949

この信頼区間は R の既存の関数 *t.test* を用いても容易に計算できる。本来仮説検定用の関数で信頼区間が計算できる仕組みは次の節で述べる。

3.2.2 R の関数 “t.test” を用いる場合

```
t.test(y$Maruaji, conf.level = 0.95)
```

One Sample t-test

```
data: y$Maruaji
t = 68.582, df = 14, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.2202175 0.2344360
sample estimates:
mean of x
0.2273268
```

Lecture11-HW1 それぞれの種に対して, “t.test” を用いて, 分散未知の仮定の下で平均パラメータ μ に対する 90%, 95%, 99% 信頼区間を求めなさい。

4 1 変数の正規分布に対する仮説検定

4.1 仮説検定の復習

Lecture 09 の資料をまずは復習してください。

ここではマルアジだけに注目し, マルアジのプロポーションが 0.22 かどうかを検定する, というふうには問題を簡略化させてみよう (0.22 という値に特に意味はなく, 説明の都合上である)。仮説検定では帰無仮説と対立仮説という 2 つの種類の仮説を考える。

マルアジの例で 2 つの仮説を表現すると下記の通り. ただし, $\mu_0 = 0.22$ である.

帰無仮説 $H_0: \mu = \mu_0$

対立仮説 $H_1: \mu \neq \mu_0$

4.2 R を用いた μ に対する仮説検定の実施 (分散未知)

4.2.1 自分で式をつくって計算する場合

```
#Maruaji
ybar <- mean(y$Maruaji)
sigma.est <- sd(y$Maruaji)
n <- length(y$Maruaji)
mu0 <- 0.22
Tstat <- (ybar-mu0)/(sigma.est/sqrt(n))
Tstat

[1] 2.210396

alpha <- 0.05
if(abs(Tstat)<=qt(1-alpha/2,n-1)) print("H0 accepted") else print("H0 rejected")

[1] "H0 rejected"
```

4.2.2 R の関数 "t.test" を用いる場合

```
# use "t.test"
t.test(y$Maruaji, mu=mu0, alternative="two.sided", conf.level=0.95)
```

One Sample t-test

```
data: y$Maruaji
t = 2.2104, df = 14, p-value = 0.04423
alternative hypothesis: true mean is not equal to 0.22
95 percent confidence interval:
 0.2202175 0.2344360
sample estimates:
mean of x
0.2273268
```

なお, 関数 `t.test` のアウトプットを見ると, デフォルトで p 値という値と信頼区間が出力される. マルアジの例では $p = 0.04423$ であった. ここで p 値とは, その観測値において帰無仮説が棄却される最小の有意水準の

ことで, p 値が事前に定めた有意水準よりも小さければ, 帰無仮説が棄却されることになる (この場合は 0.05 よりも小さいので棄却).

Lecture11-HW2 それぞれの種に対して, 分散未知の仮定の下で平均パラメータに対して $H_0: \mu = \mu_0, H_1: \mu \neq \mu_0$ を有意水準 5% で仮説検定を行いなさい. 但し, マアジ, カイワリ, マルアジ, モロについてそれぞれ帰無仮説 $\mu_0 = 0.25, 0.45, 0.25, 0.25$ に対して検定を適用すること.

5 正規分布における 2 標本検定 (興味があれば取り組んでみて下さい)

アジ科の 4 種の形態比較をしてみる. ここでは 2 種間の比較に焦点を当てる. これを統計モデルとして次のように一般的に表現する.

$$\begin{aligned} Y_{11}, Y_{12}, \dots, Y_{1n_1} &\stackrel{iid}{\sim} N(\mu_1, \sigma_1^2) \\ Y_{21}, Y_{22}, \dots, Y_{2n_2} &\stackrel{iid}{\sim} N(\mu_2, \sigma_2^2) \end{aligned} \quad (2)$$

いま, 次のような仮説検定を考えよう.

帰無仮説 $H_0: \mu_1 - \mu_2 = d$

対立仮説 $H_1: \mu_1 - \mu_2 \neq d$

$d = 0$ のときに期待値の同一性の検定であるが, ここでは一般性を保つために, あえて差を表す定数 d を加えている. このように 2 群の比較をする際は, 期待値と併せて散らばり具合も比較をしたいと考える場合もあるだろうが, ここでは分散が等しいかどうか分からない場合も含めて, 期待値の差だけに興味があるとする. 実は, 正規分布における期待値の差の検定は古くから議論されている問題であり, 等分散が仮定できる場合の理論は, 1 標本の場合とほぼ同様に極めて自然に展開が可能である. しかしながら, 2 群において分散が等しいと自信をもって仮定できる場合は限られているのではないだろうか. 一方で, 不等分散の場合にはいささかややこしいが, R を用いる範囲では等分散であろうと不等分散であろうと労力は同じである. 等分散性の検定を行ってからどちらかの検定法を選ぶという手続きもしばしば利用される.

5.0.1 i) 等分散 $\sigma_1 = \sigma_2$ を仮定できる場合の 2 標本検定

この場合, $\bar{Y}_1, \bar{Y}_2, S_1 + S_2$ が互いに独立で, かつパラメータ (μ_1, μ_2, σ^2) の十分統計量となる. ただし,

$$\begin{aligned} \bar{Y}_i &= \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \quad \sim N(\mu_i, \sigma^2) \quad (i = 1, 2) \\ S_i &= \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 \quad \sim \sigma^2 \chi^2(n_i - 1) \quad (i = 1, 2) \end{aligned} \quad (3)$$

であり, 異なる群のデータも互いに独立であるから,

$$\begin{aligned} \bar{Y}_1 - \bar{Y}_2 &\sim N\left(\mu_1 - \mu_2, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right) \stackrel{d}{=} N\left(d, \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \sigma^2\right) \\ S_1 + S_2 &\sim \sigma^2 \chi^2(n_1 + n_2 - 2) \end{aligned} \quad (4)$$

が成り立つ。ここで上式で χ^2 分布の期待値の性質から

$$\hat{\sigma}^2 = \frac{1}{n_1 + n_2 - 2}(S_1 + S_2) \quad (5)$$

が不偏推定量であることが明らかなので、1 変数の場合と同様に統計量 T を定義し、帰無仮説の下で

$$T = \frac{\bar{Y}_1 - \bar{Y}_2 - d}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \hat{\sigma}^2}} \sim t(n_1 + n_2 - 2) \quad (6)$$

が成り立つことを利用して仮説の検定を行うことができる。たとえば採択域は $|T| \leq t_{\alpha/2}(n_1 + n_2 - 2)$ となる。

例題としてマアジとマルアジにプロポーションの差があるかどうかを検定すると以下の結果となり、差が有意であることが分かる

```
# For your computation
n1<-length(y$Maaji); y1bar <- mean(y$Maaji); S1 <- (n1-1)*var(y$Maaji)
n2<-length(y$Maruaji); y2bar <- mean(y$Maruaji); S2 <- (n2-1)*var(y$Maruaji)
df <- n1+n2-2
sigma2.est <- (S1+S2)/df
Tstat <- (y1bar-y2bar)/sqrt((1/n1+1/n2)*sigma2.est)
alpha <- 0.05
data.frame(Tstat, Threshold=qt(1-alpha/2,df), Pvalue=2*(1-pt(abs(Tstat),df)) )
```

	Tstat	Threshold	Pvalue
1	3.733184	1.984984	0.0003205096

R の検定用の関数を使う場合には、次のように計算すると同じ結果が得られる。

```
# Using function t.test
t.test(y$Maaji, y$Maruaji, mu=0, var.equal=T, alternative="both.sided")
```

```
Two Sample t-test

data: y$Maaji and y$Maruaji
t = 3.7332, df = 96, p-value = 0.0003205
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.009580294 0.031336083
sample estimates:
mean of x mean of y
0.2477849 0.2273268
```

5.0.2 ii) 不等分散 $\sigma_1 \neq \sigma_2$ の場合の 2 標本検定

等分散性を仮定できない場合にも,

$$\bar{Y}_1 - \bar{Y}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \quad (7)$$

が成り立つから, 帰無仮説の下で

$$\frac{\bar{Y}_1 - \bar{Y}_2 - d}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1) \quad (8)$$

が成り立ち, $\hat{\sigma}_i^2 = S_i/(n_i - 1)$ ($i = 1, 2$) を代入すれば等分散の時と同様に t 分布による検定ができそうに思える. しかし, このようにして作られる統計量は分散比 σ_1^2/σ_2^2 に依存する. 詳細は割愛するが, t 分布の自由度は近似的に

$$DF = [w^2/(n_1 - 1) + (1 - w)^2/(n_2 - 1)]^{-1} \text{ where } w = \frac{\hat{\sigma}_1^2/n_1}{\hat{\sigma}_1^2/n_1 + \hat{\sigma}_2^2/n_2} \quad (9)$$

で計算される. 必ずしも整数とならないため, 数表を使う際には保守的に小数点以下を切ってもよいが, R では小数値の自由度でも計算でき, `t.test` 関数もそれに対応している. このような検定を特別に Welch の検定とよぶ.

```
# For your computation
tmp1 <- var(y$Maaji)/n1
tmp2 <- var(y$Maruaji)/n2
w <- tmp1/(tmp1+tmp2)
DF <- (w^2/(n1-1)+(1-w)^2/(n2-1))^-1

Tstat <- (y1bar-y2bar)/sqrt(tmp1+tmp2)
alpha <- 0.05
data.frame(Tstat, DF, Threshold=qt(1-alpha/2,DF), Pvalue=2*(1-pt(abs(Tstat),DF)) )
```

	Tstat	DF	Threshold	Pvalue
1	5.109834	28.76471	2.045957	1.910373e-05

```
# Using t.test
t.test(y$Maaji, y$Maruaji, mu=0, var.equal=F, alternative="both.sided")
```

Welch Two Sample t-test

data: y\$Maaji and y\$Maruaji

t = 5.1098, df = 28.765, p-value = 1.91e-05

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

```
0.01226681 0.02864956
sample estimates:
mean of x mean of y
0.2477849 0.2273268
```

6 提出課題

提出物は以下の 5 つです。2 回分の課題ですので 20 点分です。

- 前回の Lecture10 の気温の図（場所も記載してください）
- 上記の Lecture11-HW1(用いたコードと結果)
- 上記の Lecture11-HW2(用いたコードと結果)
- 下記の Lecture11-HW3(用いたコードと結果)
- 下記の Lecture11-HW4(用いたコードと結果)

提出期限は 8 月 7 日 23:55 とします。前回まで同様に、1 つのワードファイルに保存してください。

Lecture11-HW3 Lecture08-HW1 の生存日数のデータを用いて平均パラメータ μ に対する 90%, 95%, 99% 信頼区間を求めなさい。下記のようにデータを読み込んで関数 t.test を用いる。

```
y <- c(1472, 1486, 1401, 1350, 1610, 1585)
```

Lecture11-HW4 Lecture09-HW1 の喫煙のデータを用いて有意水準 5% で仮説検定を行いなさい。下記のようにデータを読み込んで関数 t.test を用いる。

```
y <- c(-1, 3, ...)
```