Fisheries Population Analysis

Lecture 8 "Statistical tests”

Toshihide Kitakado

June 6, 2018

例題

カイワリを31個体ランダムにサンプリングし，そのプロポーションを測定した．これらが以下のように独立同一に正規分布に従うとするとき，　μの９５%信頼区間を求めよ．
またμ=0.2を帰無仮説とするとき，この仮説に対する両側検定の結果を求めよ．

$$Y_1, Y_2, \ldots, Y_n \sim (iid) N(\mu, \sigma^2)$$

ただし，分散は未知とする．

ところで信頼区間をどのように構成？

- 定めた水準 1-α に合うよう，観測データの確率分布に基づき区間を構成する (95%信頼水準なら α=0.05)

$$Y_1, Y_2, \ldots, Y_n \sim (iid) N(\mu, \sigma^2)$$

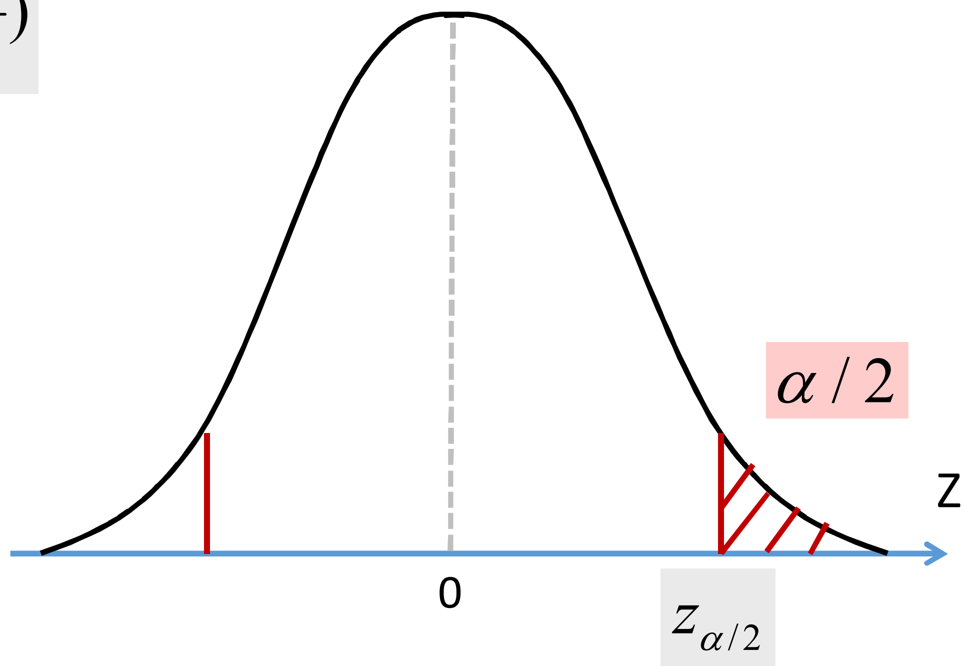
$$\bar{Y} = \frac{Y_1 + Y_2 + \cdots + Y_n}{n} \sim N(\mu, \frac{\sigma^2}{n})$$

$$\bar{Y} - \mu \sim N(0, \frac{\sigma^2}{n})$$

$$Z = \frac{\bar{Y} - \mu}{\sqrt{\sigma^2 / n}} \sim N(0,1)$$

$$P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1 - \alpha$$

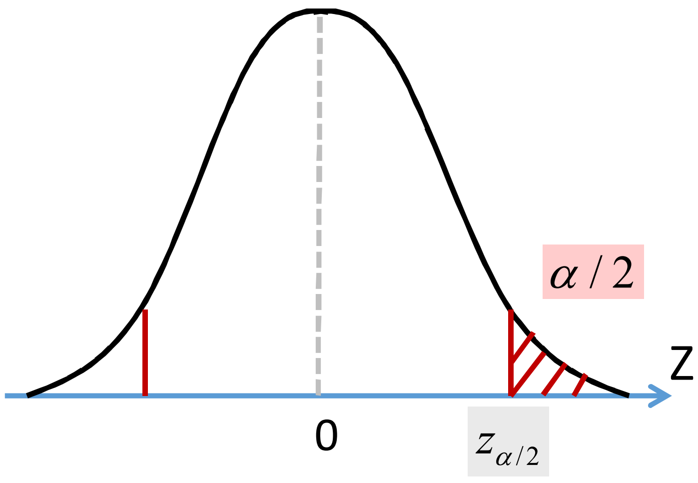
$$\bar{Y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{Y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$



分散が既知の場合と未知の場合の違いは？

$$Z=\frac{\bar{Y}-\mu}{\sqrt{\sigma^2/n}}=\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$$

$$P(-z_{\alpha/2}\le Z\le z_{\alpha/2})=1-\alpha$$



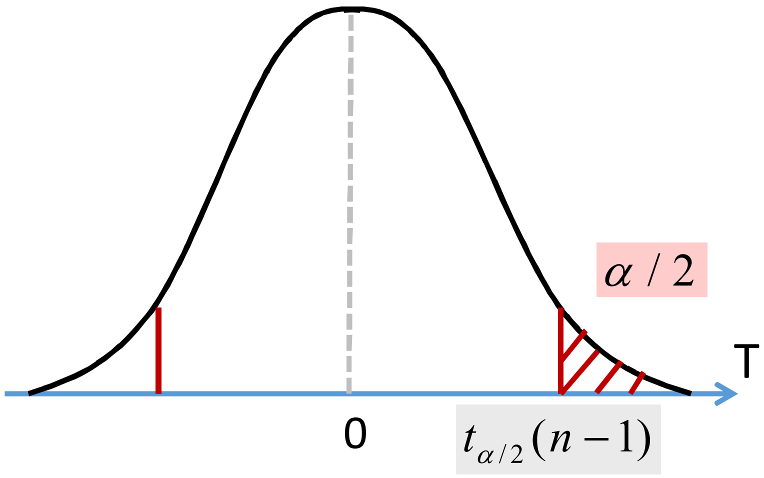
残念ながら，いま上記の式は使えない
(なぜなら，分散が未知だから！）
そこで，分散の推定値を代入するが，正規分布の仮定が崩れる

$$\hat{\sigma}^2=\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2$$

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

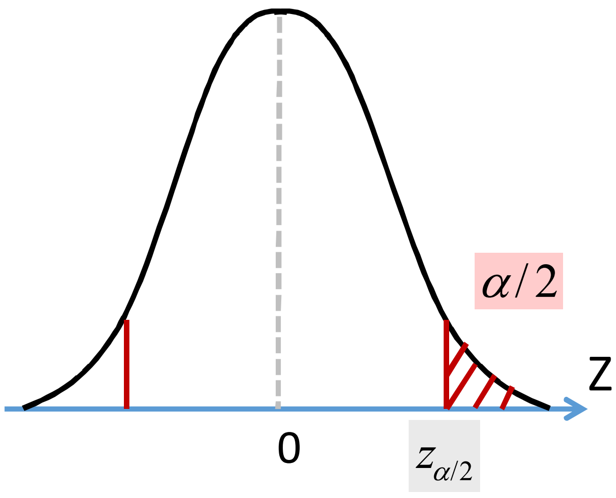
$$P(-t_{\alpha/2}\le T\le t_{\alpha/2})=1-\alpha$$

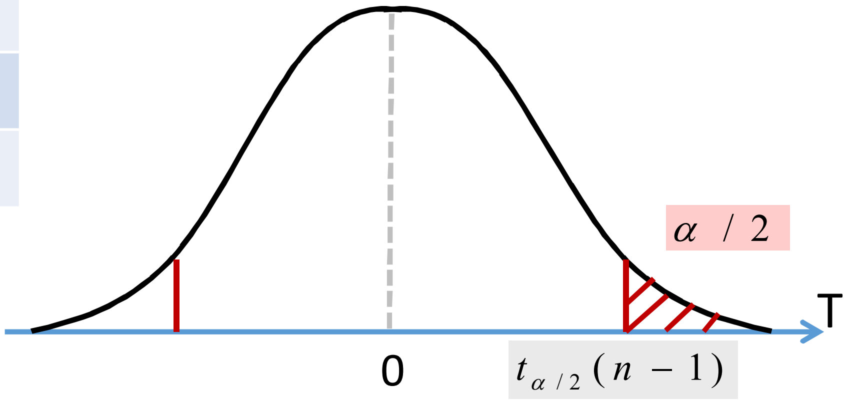
自由度n-1
のt分布



分散が既知の場合と未知の場合の違いは？

	標本数n	自由度	α = 0.05
$Z_{\alpha/2}$	-	-	1.96
$t\ \alpha/2(n-1)$	2	1	12.706
	3	2	4.303
	4	3	3.182
	5	4	2.276
	9	8	2.306
	10	9	2.262
	∞	∞	1.96

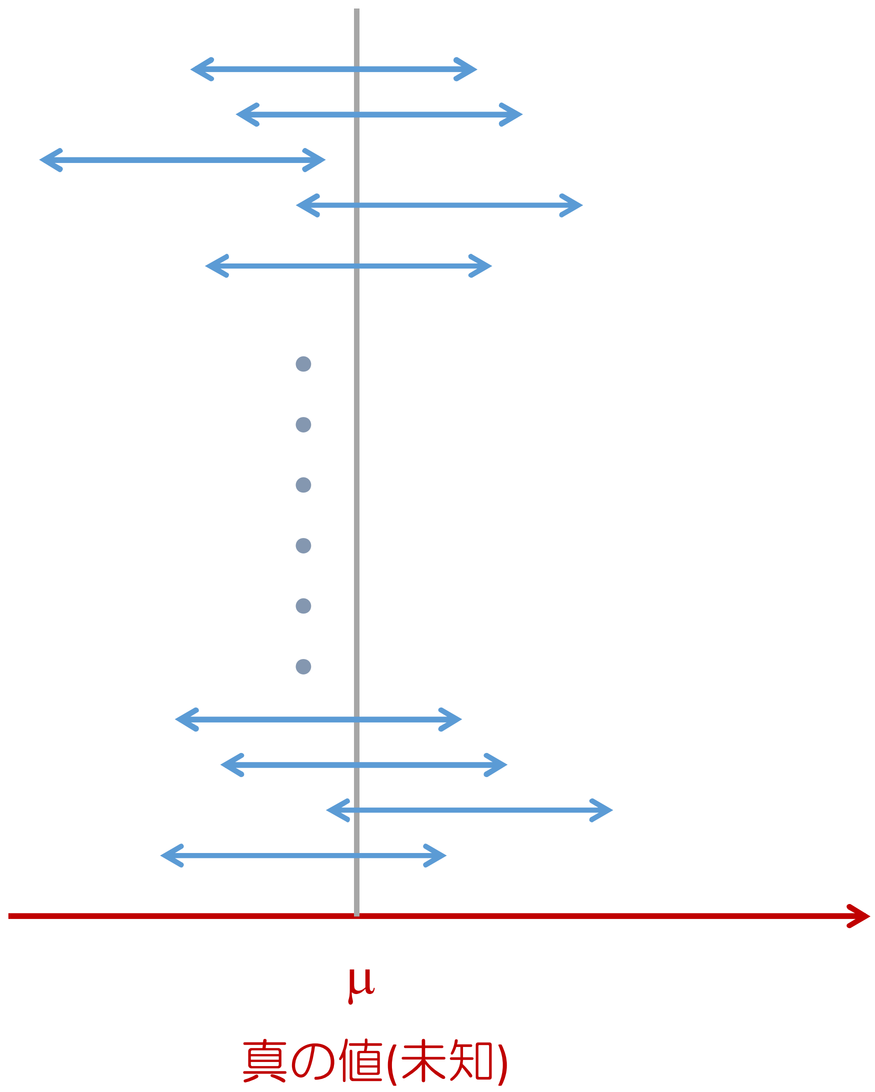




分散が未知のため，不確実性が大きくなる．よって自動的に安全を見越した広めの区間となる

信頼区間の考え方

- μの値は何か分からないが，ある値であるとする
- 信頼水準を95%と定める
- サンプル数 nのデータに対して信頼区間を計算する
- 仮にそういう操作を何度も繰り返したとするとき(例えば100回)，そのうち95%の頻度で真のμの値を含む
- データの解析者は１セットしかデータをもっていないので，そのデータから構成した信頼区間が真の値を含んでいるどうかは分からない(確率的にしか分からない)



仮説検定の流れ

1. 帰無仮説と対立仮説を設定
2. 観測データの確率分布を定義し，それぞれの仮説と対応させる
3. 検定の有意水準(第1種の過誤の確率)を設定(α=0.05など)
4. 帰無仮説が正しいと仮定する
5. 有意水準に合わせて帰無仮説の採択域と棄却域を設定
  ［帰無仮説が正しいと仮定してある統計量を定義し，その確率分布に応じて採択域と棄却域を設定］
6. 帰無仮説を採択する (疑わしきは罰せず)か, あるいは（積極的に）帰無仮説を棄却し対立仮説

仮説検定：ポテチ問題

$$Y_1, Y_2, \ldots, Y_n \sim (iid) N(\mu, \sigma^2)$$

帰無仮説 (H_0) μ=100

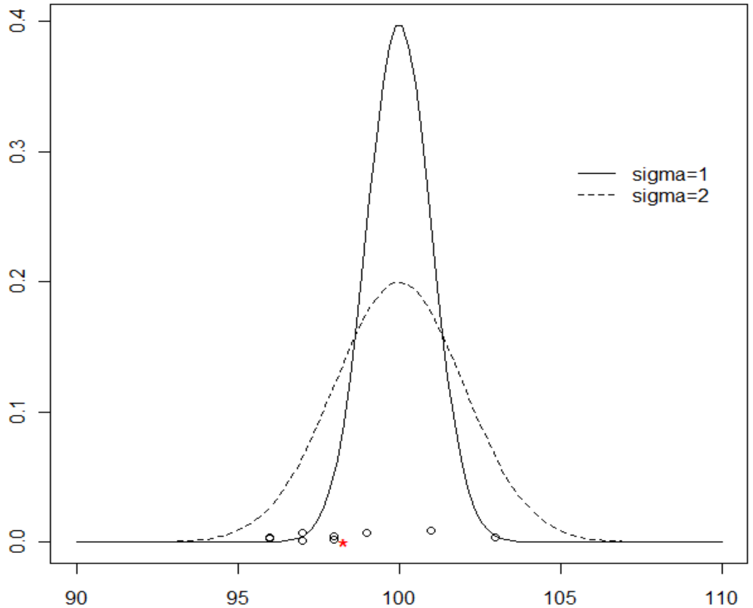
対立仮説 (H_1) μ≠100

いま，10袋の計測結果が以下のとおりであったとする

Y=(98, 96, 96, 101, 103, 99, 98, 97, 97, 98)

この平均値は98.3であり，この値が100に値に近ければH0を採択し，100から離れていればH0を棄却する

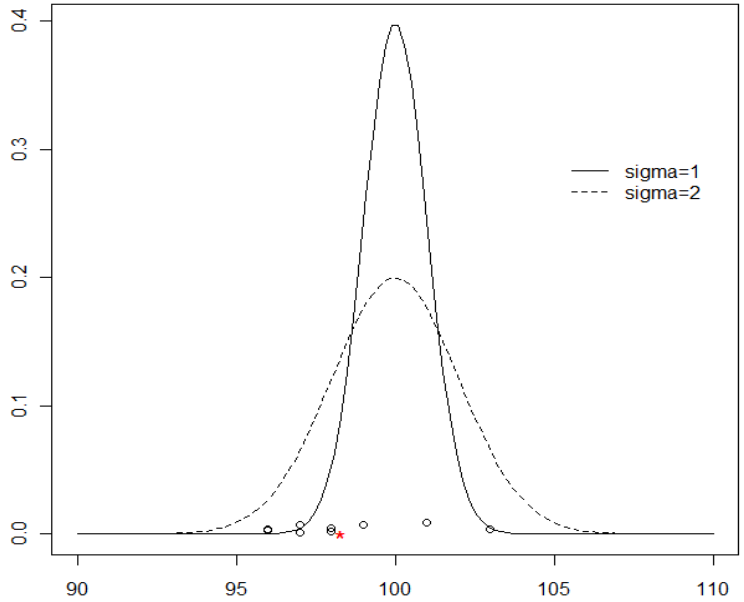
では，どれくらい離れていると棄却に帰無仮説を棄却すべきか？



ポテチ問題

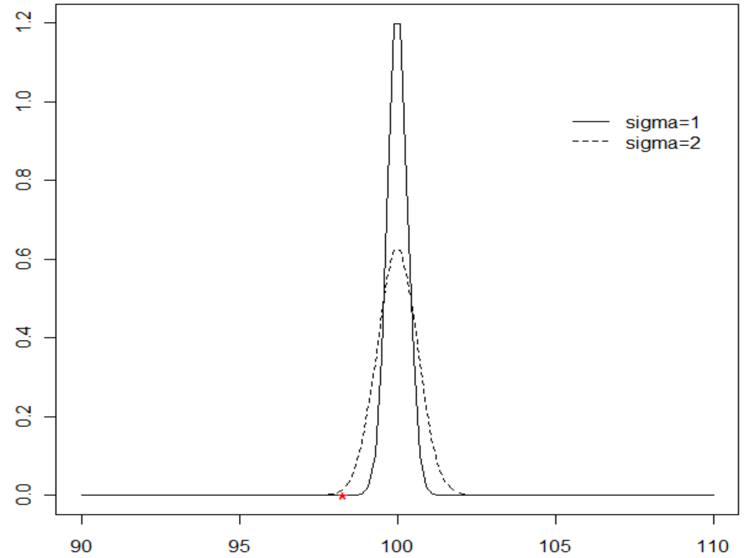
帰無仮説 (H_0) μ=100 が正しいとする

$$Y_1, Y_2, \ldots, Y_n \sim (iid) N(\mu_0, \sigma^2)$$



帰無仮説 (H_0) μ=100 が正しいとして標本平均の値に注目すると，随分可能性の低い値であることが分かる

$$\bar{Y} \sim N(\mu_0, \frac{\sigma^2}{n})$$



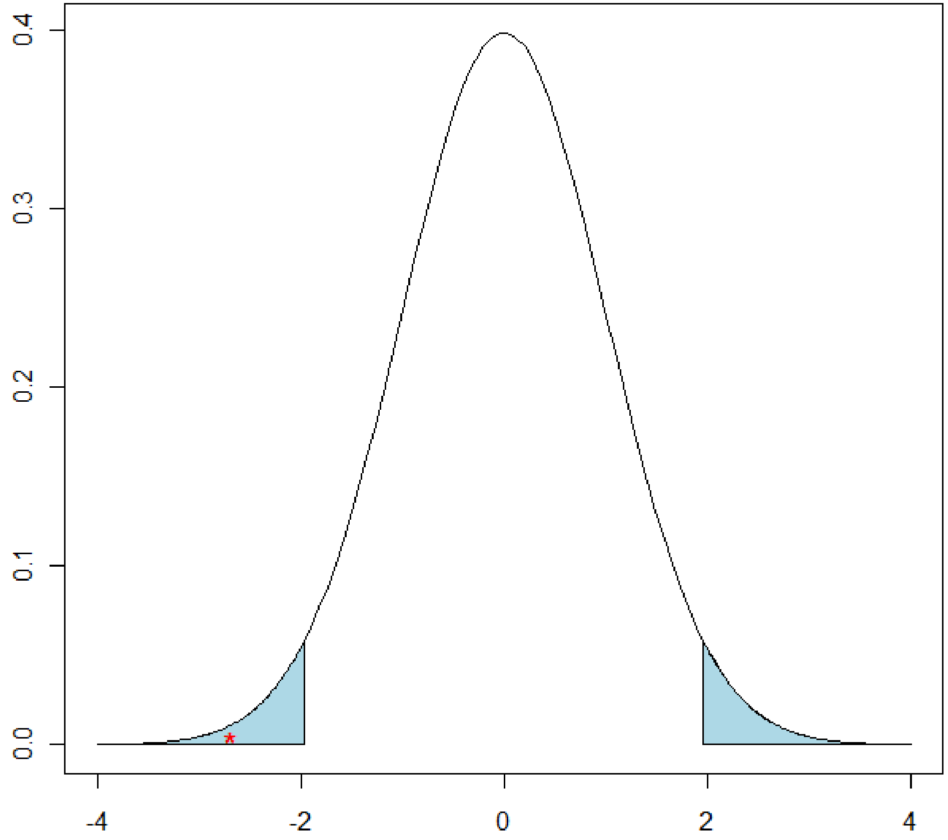
ポテチ問題

σ=2と仮定し，さらに標準化してみると

$$Z = \frac{\bar{Y} - \mu_0}{\sqrt{\sigma^2 / n}} \sim N(0,1)$$

下側2.5%の領域域に入っていることが分かる

すなわち，帰無仮説を仮定すると，かなり可能性の低いことが起きていることになる

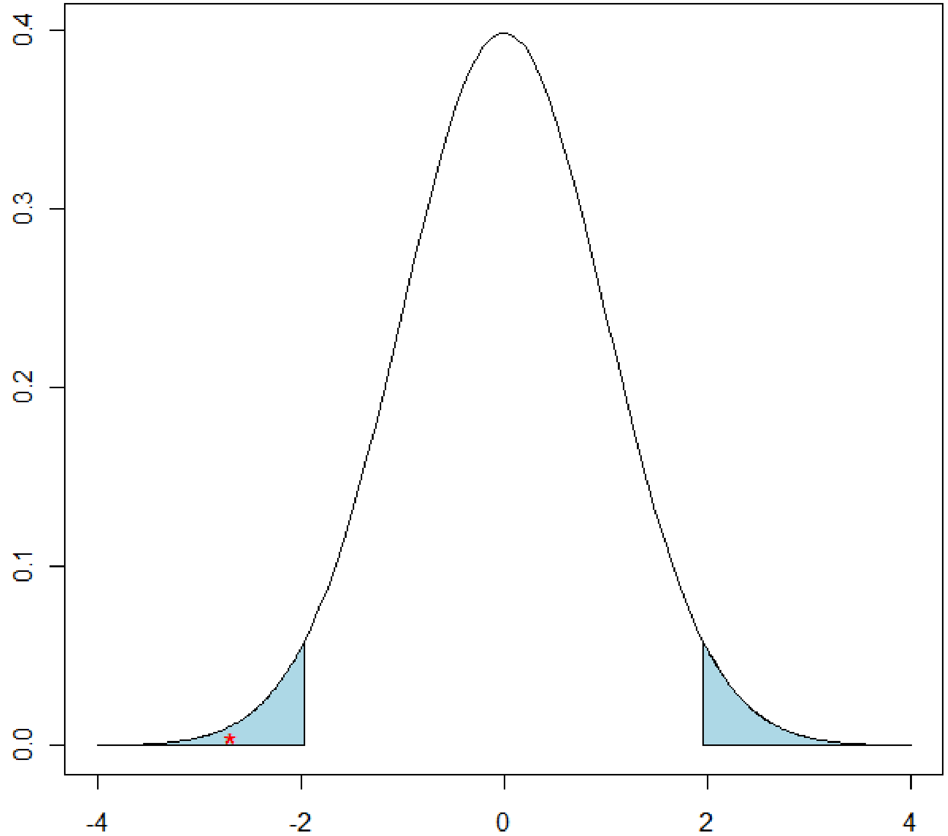


採択域と棄却域の設定

$$Z = \frac{\bar{Y} - \mu_0}{\sqrt{\sigma^2 / n}} \sim N(0,1)$$

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$

$$P(Z < -z_{\alpha/2},\ \ z_{\alpha/2} < Z) = \alpha$$



有意水準αの採択域

$$-z_{\alpha/2} \leq Z \leq z_{\alpha/2}$$

念のためZの値が採択域に含まれるかどうか確認してみて下さい

ポテチ問題(分散未知のとき)

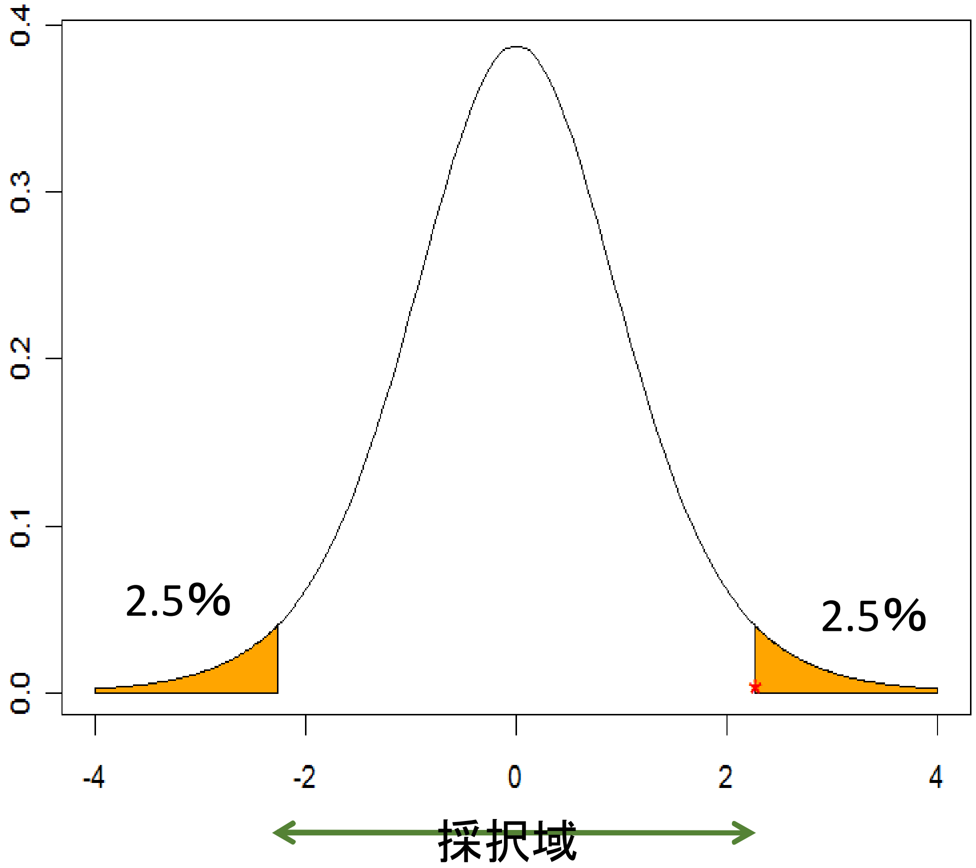
$$Y_1, Y_2, \ldots, Y_{10} \sim (iid) N(\mu, \sigma^2)$$

$$Z = \sqrt{n}\frac{\bar{Y} - \mu}{\sigma} \sim N(0,1)$$

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

$$\hat{\sigma} = 2.214$$

$$T = -2.429$$



-2.262 ≦ T ≦ 2.262

ポテチ問題 (Rを使うと)

```
x <- c(98, 96, 96, 101, 103, 99, 98, 97, 97, 98)
t.test(x, mu=100)

        One Sample t-test

data:  c(98, 96, 96, 101, 103, 99, 98, 97, 97, 98)
t = -2.4286, df = 9, p-value = 0.03807
alternative hypothesis: true mean is not equal to 100
95 percent confidence interval:
 96.71649 99.88351
sample estimates:
mean of x
     98.3
```

アジのデータ解析： t 検定

```
# One sample test (H0:mu=0.2)
Tstat1 <- sqrt(Ns)*(Mean-0.2)/SD
Tstat1

lapply(y, t.test, mu=0.2)
res.t.test <- lapply(y, t.test, mu=0.2)

Tstat2 <- Pvalue <- numeric(4)
CI <- array(NA, c(4,2))

for(i in 1:4){
 Tstat2[i] <- res.t.test[[i]]$statistic
 Pvalue[i] <- res.t.test[[i]]$p.value
 CI[i, ] <- res.t.test[[i]]$conf.int
}
data.frame(Nsample=Ns, Mean, SD, Tstat1, Tstat2, Pvalue, CI)
```

```
alternative = c("two.sided", "less", "greater")
```

3. 仮説検定(2)

2標本の検定

```
# Two sample test
t.test(y$Kaiwari,y$Maaji)
t.test(y$Kaiwari,y$Maaji, var.equal=T)

# pairwise test
pairwise.t.test(Prop,Species, pool.sd = TRUE)
pairwise.t.test(Prop,Species, pool.sd = FALSE)
pairwise.t.test(Prop,Species, pool.sd = FALSE, p.adj = "bonf")
```

分散分析

```
# Analysis of variance (分散一定)

res.lm <- lm(Prop~Species)
summary(res.lm)

res.lm <- lm(Prop~Species-1)   #切片なし
summary(res.lm)

res.anova1 <- anova(res.lm)
res.anova1

# Analysis of variance (分散不均一)

res.bartlett <- bartlett.test(Prop~Species)      #分散の均一性の検定
res.bartlett

res.anova2 <- oneway.test(Prop~Species, var.equal = FALSE)
res.anova2
```

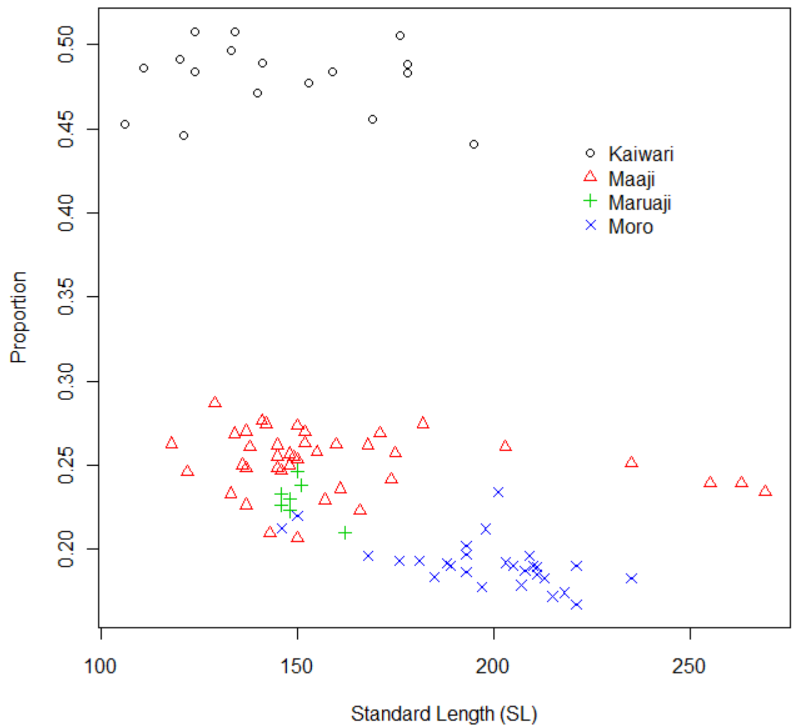
共分散分析

Analysis of covariance

```
plot(Prop~SL, pch=SP, col=SP, xlab="Standard Length (SL)",
  ylab="Proportion")
legend(220, 0.45, legend=levels(Species), pch=1:4, col=1:4, bty="n")

res.lm2 <- lm(Prop~Species + SL - 1)
res.ancova1 <- anova(res.lm2)
res.ancova1

res.lm3 <- lm(Prop~Species*(SL-1))
res.ancova2 <- anova(res.lm3)
res.ancova2
```

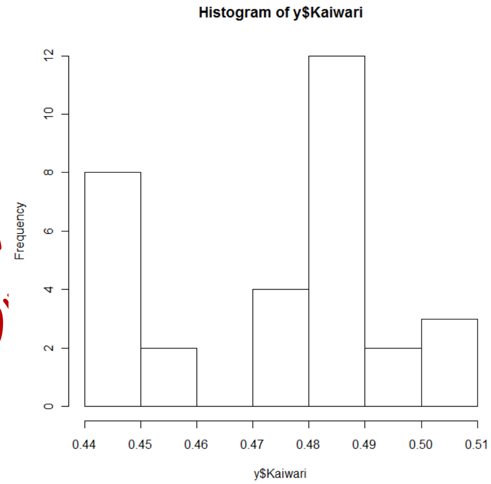


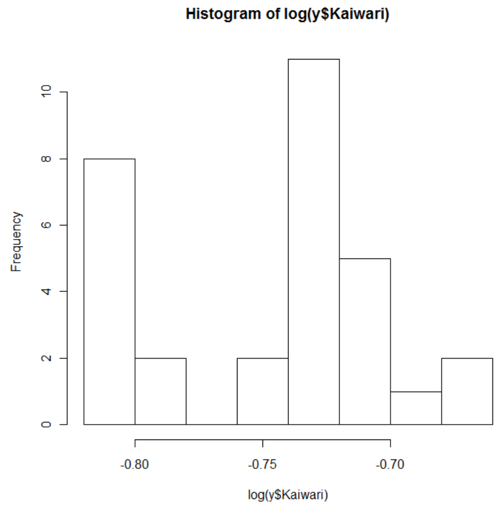
ノンパラメトリック検定

hist(y$Kaiwari)

hist(log(y$Kaiwari))

●データが正規分布に従うかどうか自信が持てないことがある





●正規分布に従っていないにデータに対して，正規分布を想定した仮説検定を行うと適切な有意水準が保たれない可能性

（標本数が大きい時には中心極限定理がはたらくため，気にしなくてもよいが）

ノンパラメトリック検定

① １標本の平均の検定 (Wilcoxonの符号付順位和検定)
wilcox.test(y$Maaji, mu=0.4)

② 2標本の平均の検定 (Mann-Whitney検定, Wilcoxon U検定)
wilcox.test(y$Maaji, y$Kaiwari)

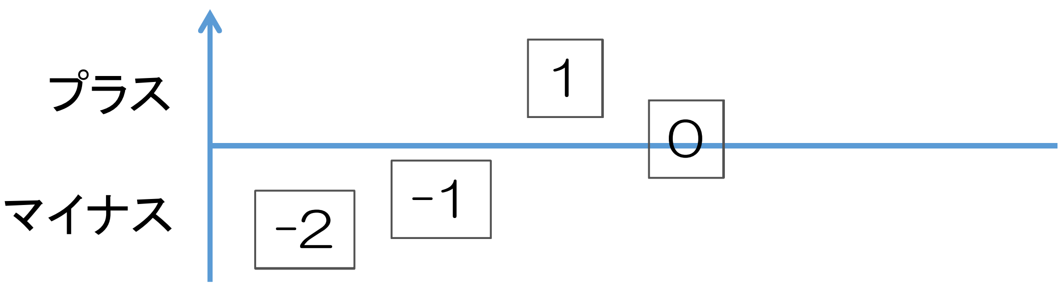
① 対応のある2標本の平均の検定 (Wilcoxonの符号付順位検定)
"wilcox.test"(x, y, paired=T)

ノンパラメトリック検定

① 1標本の平均の検定（Wilcoxonの符号付順位検定）
③ 対応のある2標本の平均の検定（Wilcoxonの符号付順位検定）
上記の二つはまったく同じ考え方で対応

例「1標本の平均の検定」の場合，データから帰無仮説の平均の値をひく

ガザミ生産日数の帰無仮説＝20日
観測した生産日数=(18, 19, 21, 20)
観測した生産日数−帰無仮説の値=(−2, −1, 1, 0)
絶対値の順位=(3, 1.5, 1.5, −)



少ない方の符号の順位を足すと 1.5

2標本の検定（対応がない場合）

Mann-Whitney検定，WilcoxonのU検定

例題（山田&北田より改題）

ガザミ生産日数（水温違い）

```
yA <- c(18,19,20,21)
yB <- c(20,21,22,21,23)
wilcox.test(yA,  yB, exact=T)

wilcox.test(y$Maaji, y$Kaiwari)
```

ノンパラメトリック検定

例えば，平均が5で自由度が１のｔ分布（コーシー分布）に従う10個のデータに対して，有意水準５％でt検定を実施すると，帰無仮説が正しくても，有意水準が２％と厳しくなっていることが分かる（水準と合っていない）

```
Nit <- 10000
Ns <- 10 ### Ns <- 100, Ns <- 1000
mu.true <- 5
Pvalue <- numeric(Nit)
for(i in 1:Nit){
 y <- mu.true + rt(Ns, df=1)
 Pvalue[i] <- t.test(y, mu= mu.true)$p.value
}
mean(Pvalue < 0.05)
```

ノンパラメトリック検定

ではWilcoxon検定の場合は？

```
Nit <- 10000
Ns <- 10 ### Ns <- 100, Ns <- 1000
mu.true <- 5
Pvalue <- numeric(Nit)
for(i in 1:Nit){
 y <-  y <- mu.true + rt(Ns, df=1)
 Pvalue[i] <- wilcox.test(y, mu= mu.true, exact=T)$p.value
}
mean(Pvalue < 0.05)
```

ノンパラ検定は外れ値に強い

ノンパラメトリック検定

ではノンパラメトリック検定ならいつも大丈夫なのか？？？
平均が5の指数分布に従う10個のデータに対して，有意水準５％でt検定を実施すると，帰無仮説が正しくても9%くらいの割合で棄却してしまうことが分かる（通常の t 検定もほぼ同様）

```
Nit <- 10000
Ns <- 10 ### Ns <- 100, Ns <- 1000
mu.true <- 5
Pvalue <- numeric(Nit)
for(i in 1:Nit){
 y <- rexp(Ns, rate=1/mu.true)
 Pvalue[i] <- wilcox.test(y, mu= mu.true, exact=T)$p.value
}
mean(Pvalue < 0.05)
```

この検定では平均値について対称であることが前提！