```
#####################################################
# 第八回「検定」
#####################################################


#####################################################
# 正規性検定
# ワールドシリーズでのアメリカンリーグ（AL）と
# ナショナルリーグ（NL）の勝敗とホームとアウェイでの
# 勝率に関するデータ
x <- read.csv("worldser.csv") #データを読み込み
str(x)
x1 <- x$natlhome #NLのホーム勝率
x2 <- x$natlaway #NLのアウェイ勝率

# Shapiro-Wilk検定で正規性検定
shapiro.test(x1)  #正規性は棄却されず
shapiro.test(x2)  #正規性は棄却されず

# K-S検定もやってみる
ks.test(x1,"pnorm",mean=mean(x1),sd=sd(x1))
ks.test(x2,"pnorm",mean=mean(x2),sd=sd(x2))

# いずれにせよ棄却されなかった.
# 正規性を仮定してt検定を適用してみる.
# t検定をしてみてホームとアウェイで違いがあるかを検定
#####################################################
# t検定：平均値の同一性検定
t.test(x1, x2) #p-value= 6.3e-11
# 平均値が等しいという仮説は棄却して良いだろう.
# ホームとアウェイでで有意な差があった.

t.test(x1, x2, var.equal=T) #p-value= 5.8e-11

#####################################################
# ノンパラメトリック検定
# Wilcoxonの順位和検定
wilcox.test(x1,x2) #やはり棄却


#####################################################
# chi^2検定
#####################################################
# 適合度検定
# 等確率性の検定
chisq.test(c(8, 12, 10, 9, 5, 6)) #棄却されず
chisq.test(c(80,120,100,90,50,60)) #同じ比率でもサンプル数が違えば棄却される

chisq.test(c(30,20,7,4), p=c(4, 3, 2, 1)/10) #こちらが指定した確率と等しいか検定
#棄却されず

#####################################################
# 独立性検定
# タイタニック号に乗船していた2201人に関する,
# 性別, 年齢(大人か子供か), 乗船区分(1等, 2等, 3等, 乗組員)と
# その人の生存・死亡についてのデータ.
# Age (0：子供, 1：大人)
# Sex (0：女性, 1：男性)
# Survived (0：死亡, 1：生存)
z <- read.csv("titanic.csv")
x <- matrix(c(sum(z$Age==1 & z$Survived==1),
              sum(z$Age==1 & z$Survived==0),
              sum(z$Age==0 & z$Survived==1),
              sum(z$Age==0 & z$Survived==0)),nrow=2)
x
chisq.test(x)  #独立性棄却
x <- matrix(c(sum(z$Sex==1 & z$Survived==1),
              sum(z$Sex==1 & z$Survived==0),
              sum(z$Sex==0 & z$Survived==1),
              sum(z$Sex==0 & z$Survived==0)),nrow=2)
x
chisq.test(x)  #独立性棄却
```

```
# データ数が多いので棄却されやすい.
# 女性の生存率は比較的高かった.

#####################################################
# 分散分析(analysis of variance, ANOVA, ANalysis Of VAriances)
# 因子(factor): 観測値に影響を及ぼすと思われる複数の要因
# 因子の水準(level): 因子を何段階に変える
# いろいろな因子の水準の違いによる影響を推測するための手法
#「同一の正規分布に従う互いに独立な誤差」という仮定が大前提
#####################################################
# コンクリートの水セメント比のデータ：1元配置モデルの分散分析
# 成分の砂を因子, 砂の種類を水準
# 砂の種類により水セメント比に差が生じるか否かを見る
# moisture_i = mu + aggregate_i + ε_i
#####################################################
concrete <- data.frame(
  aggregate=rep(1:5,6),                              # 砂種類
  moisture=c(551,595,639,417,563,457,580,615,449,631,450,508,511,517,522,
             731,583,573,438,613,499,633,648,415,656,632,517,677,555,679))
str(concrete)                                        # aggregate 変数は整数値とされ
ている
concrete$aggregate <- factor(concrete$aggregate) # aggregate 変数を因子に変換
str(concrete)                                        # aggregate 変数は因子とされて
いる
summary(concrete)
plot(concrete)
(fm <- lm(moisture ~ aggregate, data=concrete))  # 回帰分析
summary(fm)
coef(fm)
op <- par(mfrow=c(2,2))
plot(fm)
(avv <- anova(aov(fm)))                              # 分散分析表に集計
par(op)
#####################################################################
# 子供の靴の磨耗度：2元配置データの解析
# 子供靴メーカが, 靴底の材質と子供の2つの因子と靴底の磨耗度のデータ
# 靴底の材質は2種類(2水準)
# 子供は10人(10水準)
# これらの20通りのすべての水準組合せについて実験
# 実験の目的は靴底材質による磨耗度の差にあり, 子供による差を見ることは重要ではな
い
# その効果の推定には興味がないが, 観測値が得られる環境の影響を除くために導入せざ
るを
# 得ない因子をブロック因子と呼ぶ
# wear_{ij} = μ + material_i + boys_j + ε_{ij}
#######################################################
x <- c(13.2,14.0,8.2,8.8,10.9,11.2,14.3,14.2,10.7,11.8, # 靴の磨耗度
       6.6,6.4,9.5,9.8,10.8,11.3,8.8,9.3,13.3,13.6)
y <- c(1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2)          # 靴底材質(2水準)
z <- c(1,1,2,2,3,3,4,4,5,5,6,6,7,7,8,8,9,9,10,10)        # 子供(10水準)
boxshoes <- data.frame(wear=x,material=factor(y),boy=factor(z))
str(boxshoes)
mean(boxshoes$wear)                                  # 総平均
summary(boxshoes)
plot(boxshoes)
(fm <- lm(wear ~ material+boy,data=boxshoes))  # 線形回帰
summary(fm)
op <- par(mfrow=c(2,2))
plot(fm)
# 分散分析。material,boyともに1%有意(摩耗度に影響有り)
(av <- anova(fm))
par(op)
#################################################################
# 交互作用のある5元配置モデルの分散分析
# 化学反応速度に関するデータ
# 供給速度(F), 触媒(C), 攪拌速度(A), 温度(T), 濃度(N)
# の5因子をいずれも2水準 -1,1 に取り, ある化学反応速度
# y (単位%)を測定した結果
# 全ての組み合わせ32通りの半分16通りで実験(一部実施要因実験) 2^{5-1} 実験
# y_{ijklm} = μ + F_i + C_j + A_k + T_l + N_m + ε_{ijklm}
#################################################################
reacted <- data.frame(
```

```
  y=c(56,53,63,65,53,55,67,61,69,45,78,93,49,60,95,82), # 化学反応速度(%)
  F=factor(c(-1,1,-1,1,-1,1,-1,1,-1,1,-1,1,-1,1,-1,1)), # 供給速度因子
  C=factor(c(-1,-1,1,1,-1,-1,1,1,-1,-1,1,1,-1,-1,1,1)), # 触媒因子
  A=factor(c(-1,-1,-1,-1,1,1,1,1,-1,-1,-1,-1,1,1,1,1)), # 攪拌速度因子
  T=factor(c(-1,-1,-1,-1,-1,-1,-1,-1,1,1,1,1,1,1,1,1)), # 温度因子
  N=factor(c(1,-1,-1,1,-1,1,1,-1,-1,1,1,-1,1,-1,-1,1))) # 濃度因子
str(reacted)
summary(reacted)
mean(reacted$y)                                  # 総平均
(fm <- lm(y ~F+C+A+T+N, data=reacted))           # 線形回帰
summary(fm)
op <- par(mfrow=c(2,2))
plot(fm)
(av <- aov(fm))                                  # 分散分析
(avv <- anova(av))                               # Cは1%有意,Tは5%有意
par(op)
# 因子C,Tとその交互作用だけで検討
summary(lm(y ~ C + T + C*T, data = reacted))   # y ~ (C+T)^2 でもよい
summary(aov(y ~ C + T + C*T, data = reacted))  # 交互作用も1%有意(交互作用が認め
られる)
# 5因子間のすべての交互作用を考慮したモデルの分散分析結果
summary(aov(y~(F+C+A+T+N)^2, data=reacted))
# 変数選択
summary(step(lm(y~F+C+A+T+N,data=reacted)))     # 回帰結果から変数選択実行
summary(step(lm(y~(C+T+N)^2,data=reacted)))     # 相互作用項を含め再度変数選択実
行
#################################################################
# 4種類の油脂A,B,C,Dを使い, それぞれ6個のドーナツを調理し, 油脂吸収量を測定
# 油脂種類を効果 (4水準) として, 一元配置の分散分析を実施
# 用いた油による吸収量の差が本当にあるかどうかの検定を行う
x <- c(64,72,68,77,56,95,78,91,97,82,85,77,75,93,78,71,63,76,55,66,49,64,70,68)
y <- c(rep("A",6),rep("B",6),rep("C",6),rep("D",6))
( Donut <- data.frame(group=y,obs=x) )
str(Donut)      # 変数groupは因子とされ、内部的には1,2,3,4で表現される
summary(Donut)  # 要約
plot(Donut)     # 各因子毎に箱型図を描く
(fm <- lm(obs ~ group, data=Donut)) # 目的変数 obs, 説明変数 group で線形回帰
fm
summary(fm)
mean(Donut$obs)                  # 24のデータの総平均は73.75
fitted(fm)                       # 各データ毎の推定値
coef(fm)                         # 各水準毎の切片値
op <- par(mfrow=c(2,2))
plot(fm)                         # 回帰診断図
(av <- anova(fm))                # 回帰結果から分散分析表を作成
# 有意水準1%で油脂による吸収量の差はないという帰無仮説を棄却
summary(av)
```