2010年度 統計解析演習　Rによる統計

0　半年間でやること

毎週院生室（理学部C626）において，統計ソフトRを用いて統計的推論のシミュレーションと簡単なデータ解析を学習します．具体的には以下の内容を予定していますが，必要に応じて順番の入れ替え，内容の変更を行います．

データ解析：

- データファイルの作成と読込，記述統計（統計量，ヒストグラム）
- 2群の位置母数の比較（対応の有無，t検定，Welchのt検定，Wilcoxonの順位和検定）
- 3群以上の位置母数の比較（検定の多重性，一元配置分散分析，二元配置分散分析）
- 重回帰分析（決定係数，偏回帰係数の有意性，情報量規準）
- ロジスティック回帰分析（オッズ比，交絡因子の調整）
- 離散データ解析（カイ二乗検定，Fisherの正確検定，McNemar検定）

統計的推測のシミュレーション：

- 擬似乱数（メルセンヌ・ツイスター，コイン投げ），正規乱数（Box-Muller法，再生性）
- 推定量の良さ（平均二乗誤差，不偏性，十分統計量の直感的意味）
- 区間推定（被覆確率，信頼区間プロットによる視覚化，分散安定化変換）
- 仮説検定（有意水準，p-値の数値実験，適合度検定）
- 大数の法則と中心極限定理（一致推定量，モンテカルロ法，ヒストグラムによる視覚化）
- ブートストラップ法（バイアス補正，ブートストラップ信頼区間）

ゼミの方法および目標

データ解析では参考書に載っているデータをExcelに入力し，Rの1行コマンドで結果を出力し，出力結果の解釈が出来ることを目標にします．また，統計シミュレーションでは確率・統計Bおよび今期の確率統計特殊講義の前半で学習する統計的推測の観念についてシミュレーションの方面から理解を深めます．そのため出来るだけ確率統計特殊講義の授業には出席してください．

持参物および諸注意

このセミナーには**卒業研究のテキスト**（統計データ解析入門）および**データを保存するためのUSB**を持ってきてください．また，適宜プリントを配布しますので，必要ならそれも持参してください．なお，病気などの理由で休む場合は，**必ずメールで連絡してください．**（携帯メール：tomo-statim @ i.softbank.jp）

なお，配布プリントや演習で作成したRのソースコードは以下のWWWで公開し他のゼミ生が参照できるようにする予定です．

秋田 智之（大学院理学研究科数学専攻 確率統計講座 D3）
研究室：理学部C626
E-Mail：d082905 @ hiroshima-u.ac.jp
url：http://home.hiroshima-u.ac.jp/d082905/R.html

1　データファイルの生成と読込，記述統計

対応の有無とデータの形式

データを R などの統計ソフトで解析するには，次のような形に加工する必要があります.（便宜上最初の列は ID にすることが多い.）

表 1. データファイルの形式（左：対応あり，右：対応なし）

ID	Before	After
1	X_1	Y_1
2	X_2	Y_2
$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$
$n-1$	X_{n-1}	Y_{n-1}
n	X_n	Y_n

ID	Group	Measure.
1	1	Y_1
$\vdots$	$\vdots$	$\vdots$
n_1	1	Y_{n_1}
n_1+1	0	Y_{n_1+1}
$\vdots$	$\vdots$	$\vdots$
n	0	Y_n

問 1. Excel を起動してこのどちらかの形式で参考書のデータを入力せよ.（注意：ヘッダーは入力し，罫線などは書かないこと.）

問 2. 問 1 で作成したデータを「Data0415.csv」として保存せよ.（注意：「ファイル」から「名前を付けて保存」を選び，ファイル名 (N) を「Data0415.csv」，ファイルの種類 (T) を「CSV(カンマ区切り)∗.csv)」とする.）

データの読込と記述統計

R を起動して「ファイル」から「新しいスクリプト」を選び，次を入力してください.

データの読込.

```
Data <- read.csv(''Data0415.csv'')
```

次にこれを左ドラックで選択し，右クリックで「カーソル行または選択中の R コードを実行」を選びます．このときエラーが出なかったらコンソールで `Data` と打ち Enter キーを押して中身を確認します.

注意：エラーが生じたら，まず「Data0415.csv」を右クリックして「プロパティ」を表示させます．このとき場所が「C:￥Documents and Settings￥Administrator￥My Documents」となっていたら，￥を/に変えてコピーし，スクリプトの `Data0415.csv` の前に挿入します.

次にヒストグラムを描いたり（**データの視覚化**），平均や分散を計算（**データの要約**）してデータの全体を把握します．Data のヘッダーが ID, GROUP, SCORE である場合，SCORE のヒストグラムや統計量は次のコマンドで出力できます.

記述統計.

```
hist(Data$SCORE) #SCORE のヒストグラム
mean(Data$SCORE) #SCORE の平均
var(Data$SCORE) #SCORE の分散
```

問 3. 読み込んだデータ Data のある列のヒストグラムを書いたり，統計量を計算せよ.

2　擬似乱数

乱数とは例えばサイコロを何回も投げて出た目の列 $6, 2, 2, 4, 3, 5, \ldots$ のようなものである．これをコンピュータで擬似的に再現するのが擬似乱数であり，数値実験には必須である．

線形合同法

擬似乱数の簡単な生成法として**線形合同法**がある．これは漸化式

$$x_{n+1} = ax_n + b \mod M$$

によって乱数 $x_1, x_2, \ldots$ を発生させる方法である．($\mathrm{Mod}M$ は M で割った余り)

R の文法（for ループ）．

```
for(i in 1:100){ 処理 }      # 処理を 100 回繰り返す
```

問 4．C 言語の関数 rand は $a = 1130515245, b = 12345, M = 2^{32}$ として線形合同法により乱数を発生させている．以下をスクリプトに打ち込み実行させ，コンソールで X を表示してみたり，ヒストグラムを描いてみよ．また，サイコロで出た目にするにはどうすればよいか．

線形合同法．

```
M <- 2^(32)                    # 各パラメータの設定
a <- 1130515245; b <- 12345
X <- 1                         # 漸化式の初項の設定
for(i in 1:100){               # 以下の処理を 100 回繰り返す
X <- c(X, (a*X[i]+b)%%M)       # リスト X に (a*X[i]+b)%%M を付け加える
}                              # X[i] はリスト X の第 i 成分，%%M は M で割った余り
```

メルセンヌ・ツイスター

線形合同法は計算速度が速いが欠点も多い．松本・西村により開発されたメルセンヌ・ツイスター (MT) は速さと性質の良さを兼ねそろえた乱数発生法であり，簡単に言うと整数列の代わりに 2 進法表記したベクトル列 $\{\boldsymbol{x}_n\}$ を

$$\boldsymbol{x}_{j+n} := \boldsymbol{x}_{j+m} + \boldsymbol{x}_{j+1}B + \boldsymbol{x}_j C \quad (j = 0, 1, \ldots)$$

により生成するものである．R の擬似乱数は MT により生成されている．

問 5．R では runif(N) で $[0, 1]$ 上の一様乱数を N 個発生させることができる．それらのうち $[0, 0.5]$ にあるものを表，$[0.5, 1]$ にあるものを裏と変換すると，コイン投げが再現できる．以下を打ち込み，コンソールで coin を表示させてみよ．

コイン投げの再現．

```
p <- 0.5                                  # コインで表が出る確率
coin <- (NULL)                            # 代入の都合上，空のリストを用意
for(i in 1:20){                           # 以下の処理を 20 回繰り返す
coin <- c(coin, ifelse(runif(1)<p,1,0)) # 発生させた乱数が p 未満なら 1 を
}                                         # そうでなければ 0 をリストに追加
```

R の文法（if-else 文）.

```
ifelse(条件，処理 1，処理 2) # 条件が真なら処理 1 を，偽なら処理 2 を行う
```

Box-Muller 法

統計のシミュレーションでは正規分布に従う乱数を使うことが多く，R では rnorm(N) で正規乱数を N 個発生させることができる．Box-Muller 法は一様乱数を正規乱数に変換するもので，U_1, U_2 が独立に $[0,1]$ 上の一様分布に従っているとき

$$X_1 = \sqrt{-2\log U_1}\cdot\cos(2\pi U_2), \quad X_2 = \sqrt{-2\log U_1}\cdot\sin(2\pi U_2)$$

は独立に標準正規分布に従うことが知られている．

問 6．上から一様乱数 U_1, U_2 を $X = \sqrt{-2\log U_1}\cdot\cos(2\pi U_2)$ と変換すると X は正規乱数になる．以下を打ち込み，コンソールで NData のヒストグラムを描き，正規分布になっているか確かめよ．

Box-Muller 法.

```
Data1 <- runif(1000)       # U1 のデータを 1000 個生成
Data2 <- runif(1000)       # U2 のデータを 1000 個生成
R1 <- sqrt(-2*log(Data1)) # sqrt(-2*log(U1)) のリストを作成
R2 <- cos(2*pi*Data2)      # cos(2*pi*U2) のリストを作成
NData <- R1*R2             # R1 と R2 のリストの成分ごとの積を求める
```

注意：ヒストグラムを描いた後で curve(500*dnorm(x,0,1),col=''red'',add=T) を実行すると，正規分布の密度関数を重ね書きすることが出来る．

正規分布の再生性

確率変数 X_1, X_2 が独立にそれぞれ正規分布 $N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2)$ に従うとき，その和 $X_1 + X_2$ は正規分布 $N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$ に従う．これを**正規分布の再生性**という．

問 7.「注意」で見たように標本サイズが大きいときヒストグラムと母集団分布の確率密度関数が近いことが知られている（**グリベンゴ・カンテリの定理**）．そこで，正規乱数を 1000 個ずつ発生させたリスト NData1,NData2 の平均・分散を求め，さらにリストの和 NData1+NData2 のヒストグラムを描いたり，平均・分散を求めることによって，上の定理を確かめよ．

「正規分布の再生性」のためのデータ.

```
NData1 <- rnorm(1000,3,2)  # 平均 3, 標準偏差 2 である正規乱数を 1000 個発生
NData2 <- rnorm(1000,5,4)
NDataU <- NData1+NData2    # Data1 と Data2 のリストごとの和
```

補足：この定理から後の区間推定や検定で用いられる，標本平均についての重要な定理が従います：$X_1, X_2, \ldots, X_n$ が独立に正規分布 $N(\mu, \sigma^2)$ に従うとき

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \textbf{は正規分布 } N(\mu, \frac{\sigma^2}{n}) \textbf{ に従う}$$

是非覚えて置いてください．

3　仮説検定

データ解析では統計量やヒストグラムだけでなく，検定を同時に行うことがしばしばある．

統計量の分布と仮説検定問題

正規母集団 $N(\mu, 1)$ の母平均 μ に関する仮説検定問題

$$H_0 : \mu = 0 \quad \text{v.s.} \quad H_1 : \mu > 0$$

を考える．このとき母集団からの独立標本 $X_1, X_2, \ldots, X_n$ の標本平均 $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$ がある程度大きければ H_1 であると考えられる．

問 8．以下のコマンドを実行して H_0 が真のときの $\bar{X}$ のヒストグラムを描け．（これは**帰無分布**と呼ばれる．）このことからごく稀に $\bar{X}$ が大きな値をとることがあることを確認せよ．

H_0 の元での標本平均の分布．

```
Data <- (NULL)                          # 代入の都合上，リストを初期化
AveData <- (NULL)
N <- 10                                 # 標本サイズ N を 10 とする
for(i in 1:10000){                      # 以下の処理を 10000 回繰り返す
 Data <- rnorm(N)                       # 平均 0 の正規乱数を N 個発生させる
 Ave <- mean(Data)                      # データの平均を Ave に代入
 AveData <- c(AveData, Ave)             # 標本平均のデータセットに Ave を追加
}
hist(AveData)                           # データセットのヒストグラムを描く
```

有意水準（危険率）

問 8 から H_0 か H_1 かの判断を決める $\bar{X}$ の閾値をどの値にしても**第 1 種の過誤**，即ち H_0 が真なのに H_0 を棄却する可能性があることが分かった．通常行われる仮設検定はこれが起きる確率を有意水準と呼ばれる小さな値 α で抑えるようにする．

問 9．前項の仮説検定問題において，有意水準 5% の検定方式は「$\bar{X} > 1.64\sqrt{\sigma^2/n}$ のとき H_0 を棄却」で与えられる．この第 1 種の過誤を起こす確率を求めてみよう．

第 1 種の過誤の確率．

```
Data <- (NULL)                          # リストの初期化
B <- 10000                              # シミュレーション回数
C <- 0                                  # カウンターに 0 を代入
N <- 20                                 # 標本サイズ
for(i in 1:B){                          # 以下を B 回繰り返す
 Data <- rnorm(N)                       # 正規乱数を N 個発生
 if(mean(Data)>1.64/sqrt(N))            # もしデータの平均が 1.64/sqrt(N) より
   C <- C+1                             # 大きければ，カウンターに 1 加える
}
C/B                                     # 間違いが起きた割合を出力
```

対立分布と検出力（検定力, Power）

前項では H_0 が真であるとき，H_0 を棄却する確率を求めた．今度は H_1 が正しいときに H_0 を棄却する確率を求める．この確率を**検出力**という．

問 10．問 9 のソースの 6 行目の `rnorm(N)` を `rnorm(N)+0.3` とすると，$\mu = 0.3$ であるときに H_0 を棄却する確率が求められる．真の母平均 μ や標本サイズ `N` を変更して検出力がどのように変わるかを調べよ．

検出力曲線

前項の真の母平均 μ を変化したときの検出力をグラフにしたものを表示させることを考える．これは**検出力曲線**と呼ばれ，検出したい差に対して検出力の値を求めることができる．

問 11．以下のコマンドを打ち，検出力曲線を表示させ，真の母平均 μ が 0 から離れるほど検出力が上がることを確認せよ．

検出力曲線.

```
Data <- (NULL)                  # リストの初期化
x <- (1:10)*0.05-0.05           # x の目盛り（折れ線グラフ用）
y <- (NULL)
B <- 1000                       # シミュレーション回数
N <- 100                        # 標本サイズ
for(j in 1:10){                 # 以下を j=1,2,..,10 に対して行う
 C <- 0                         # カウンターに 0 を代入
 for(i in 1:B){                 # 以下を B 回繰り返す
  Data <- rnorm(N)+0.05*(j-1)   # 母平均が 0.05*(j-1) のデータを N 個発生
  if(mean(Data)>1.64/sqrt(N))   # 標本平均が 1.64/sqrt(N) より大きければ
    C <- C+1                    # カウンターに 1 加える
 }
 y <- c(y, C/B)                 # y に H0 棄却の割合（検出力）を追加
}
plot(x,y,type="l")              # y の折れ線グラフを描く
```

p-値と仮説検定

前項までは $P(\bar{X} \geq z_{0.05}|H_0) = 0.05$ となる $z_{0.05}$ を求めて，得られた標本の標本平均 $m = \bar{x}$ が $m \geq z_{0.05}$ のとき H_0 を棄却するというものであった．p-値とは確率 $P(\bar{X} \geq m|H_0)$ のことで，p-値が 0.05 以下なら有意水準 5%で差があると言える．

問 12．本節の仮説検定のために 20 個の標本を抽出し，標本平均が 0.35 であるときに p-値を以下の 2 つの方法で求めよ．

(1) シミュレーションで求める．問 9 の `1.64/sqrt(N)` を `0.35` に変更すればよい．

(2) $\sqrt{n}\bar{X}$ が正規分布 $N(0,1)$ に従うので $\sqrt{n}\bar{x}$ の上側確率を求めても良い．R では累積確率が `pnorm` で求まるので `1-pnorm(sqrt(20)*0.35)` で求められる．

注意：第 1 種の過誤を起こす確率が α より小さいとき**保守的**と呼ばれる．

4　2群の位置母数の比較（その１）

医学，心理学，スポーツ科学など様々な分野の研究では患者群と対照群，使用前と使用後のような2群の比較，およびその差に関する検定がよく行われている．

箱髭図による2群の比較

最初に2群を箱髭図で比較しよう．Rでは `boxplot(データセット名)` で箱髭図を出力できるが，このときデータセットを複数指定し，箱髭図を並べて出力させることができる（平行箱髭図）．そのため箱髭図は2群以上の比較において，データを記述するのに適している．

問 13．§1の問2で作成したデータファイルについて，箱髭図によって2群を比較しよう．§1の表1の対応のあるデータをData1，対応のないデータをData2とするとき，以下のコマンドで箱髭図を出力できる．問2で作成したデータファイルをRに読み込み（§1参照），以下のファイル名やヘッダー名を読み替えて箱髭図を出力させ，2群に差がありそうか考えよ．

2群の箱髭図.

```
boxplot(Data1$Before, Data1$After)  # 対応のあるデータ Data1 の箱髭図
```

注意：対応のない2群のデータData2は共に `Score`[*注] の列にあるためそのままでは2群全体の箱髭図を出力されてしまう．`Data2$Score[Data$Group==1]` とすれば各群のデータのみのベクトルができるので `boxplot` により2群それぞれの箱髭図を出力することが出来る．
([*注] 表1の列名 `Measure.` から変更)

対応のない2群の有意差検定の種類

前項の箱髭図で2群の差がありそうなとき，その根拠を言うにはどうすればよいであろうか．このとき箱髭図から分かるように，一方のデータが他方のデータ全てを上回ることは稀である．そのため通常は平均値や中央値といった位置母数の比較が行われる．対応のない2群の位置母数の検定方法として(Studentの)t-検定，Welchのt-検定，Wilcoxonの順位和検定があり，どの手法を用いるかは図1より選択できる．

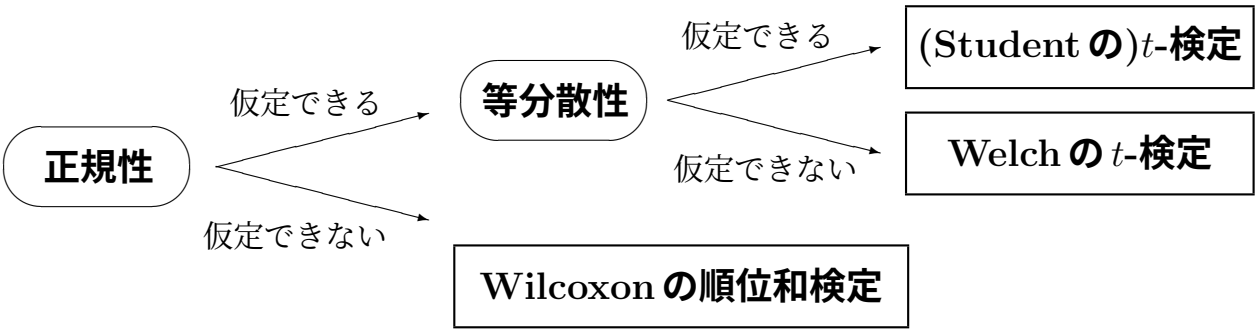

図1. 対応のない2群の位置母数に関する検定手法の選び方

問 14．データが正規分布に従っているかの確認で，正規確率紙へデータをプロットし（QQプロット）直線状に並べば正規分布，そうでなければ正規分布でないと判断する方法がある．`rnorm(100)`, `rchisq(100,2)` により正規乱数と自由度2のカイ2乗乱数の100個のデータセットを発生させ，それらのQQプロットを描くことにより，そのことを確認せよ．

正規確率紙へのプロット.

```
Data <- rchisq(100,2)          # Data に自由度 2 のカイ 2 乗乱数 100 個を代入
qqnorm(Data)                   # Data の QQ プロットを表示
```

Student の t-検定

等分散である 2 群の正規母集団 $N(\mu_1, \sigma^2), N(\mu_2, \sigma^2)$ の平均に関する仮説検定問題

$$H_0 : \mu_1 = \mu_2 \quad \text{v.s.} \quad H_1 : \mu_1 \neq \mu_2$$

を考える．それぞれの標本平均の差 $\bar{Y}_1 - \bar{Y}_2$ が 0 から離れていれば H_1 であると判断できる．

問 15．等分散性の判定には F-検定が用いられる．まず，コマンド var.test により F-検定を行え．そして出力の p-value が 0.20 以上であれば等分散であるとみなして Student の t-検定を，そうでなければ Welch の t-検定を行い，p-value から 2 群の平均に差があるかどうかの判断を下せ.（以下のコマンドは Data$を省略しているので注意せよ.）

F-検定.

```
var.test(Score[Group==1],Score[Group==0])
```

Student の t-検定（等分散性を仮定）.

```
t.test(Score[Group==1],Score[Group==0],var.equal=T)
```

Welch の t-検定（等分散性を仮定しない）.

```
t.test(Score[Group==1],Score[Group==0],var.equal=F)
```

注意：慣習的に F-検定の有意水準は t-検定の有意水準 α の 4 倍程度に定められる．例えば t-検定の有意水準が 1% のとき F-検定の有意水準は 4% である．

Wilcoxon の順位和検定（ノンパラメトリック検定）

正規性の仮定が成り立たないときは，2 群の平均を比較することは適切でないことがある．そのようなときは平均値ではなく，2 群の中央値が等しいかどうかの検定を行う．

問 16．箱髭図の箱の中の線は中央値を表している．次のコマンドにより 2 群それぞれの中央値を出力せよ．

記述統計（中央値）.

```
median(Score[Group==1])                # 1 群のデータの中央値を出力
```

注意：summary(Data2$Score[Group==1]) とすれば，中央値などが一度に出力される．

問 17．上で用いたデータを Wilcoxon の順位和検定で検定せよ．また，この際どちらの方が p-値が小さいかを比較せよ.（正規性が成り立つときは検出力が劣る）

Wilcoxon の順位和検定（Mann-Whitney 検定）.

```
wilcox.test(Score[Group==1],Score[Group==0])
```

注意：対応のあるデータの有意差検定については後で扱うことにする．

5　推定量の良さ

未知母数 θ に対していくつもの推定量が考えられるとき，それらの推定量の良さを不偏性，一致性，平均二乗誤差などの観点から比較しよう．

不偏性

推定量が満たしたほうが良いとされる性質で代表的なものに**不偏性**がある．これは θ の推定量 $\hat{\theta}$ に対して $E(\hat{\theta}) = \theta$ が成り立つことである．これは「推定値が母数よりも大きいことも小さいこともあるが，長い目で見るとどちらにも偏っていない」とみなされる．

問 18．分散 σ^2 の不偏推定量 V および最尤推定量 $\hat{\sigma}^2$ は

$$V = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad \left(\hat{\sigma}^2 = \frac{n-1}{n}V\right)$$

である．R では var(X) でデータセット X の不偏分散が得られる．次を実行し，不偏分散 V の分布のヒストグラムや平均を計算し，不偏性を確かめよ．また標本分散 $\hat{\sigma}^2$ が不偏でないことも確かめよ．

不偏分散の分布.

```
X <- (NULL); V <- (NULL)          # リスト X,V の初期化
N <- 20 ; B <- 10000              # 標本サイズとシミュレーション回数
for(i in 1:B){                    # 以下を B 回繰り返す
 X <- rnorm(N)                    # X に正規乱数を N 個代入
 V <- c(V,var(X))                 # 分散のリスト V にデータセット X の不偏分散
}                                 # を追加
```

一致性

標本 $X_1, X_2, \ldots, X_n$ による推定量を $\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \ldots, X_n)$ で表すことにする．$\hat{\theta}_n \xrightarrow{p} \theta\,(n \to \infty)$ を満たすとき $\hat{\theta}_n$ は**一致性**を持つという．このとき「標本サイズが大きいとき，推定量と母数はほぼ等しい」が成り立っていると考えて差し支えない．

問 19．大数の法則から標本平均は母平均の一致推定量であることが分かる．次のコマンドにより n が増加するごとの標本平均をプロットし，母平均 0 に近づいているかどうか調べよ．

標本平均の挙動.

```
X <- (NULL)                       # データセットの初期化
x <- (1:100)*10                   # x の目盛り（折れ線グラフ用）
y <- (NULL)                       # 標本サイズ n のときの標本平均のリスト用
for(i in 1:100){                  # 以下を 100 回繰り返す
 X <- c(X,rnorm(10))              # 10 個の正規乱数を X に追加
 y <- c(y,mean(X))                # X の標本平均をリスト y に追加
}
plot(x,y,type="l")                # y の折れ線グラフを描く
abline(h=0)                       # 比較のために y=0 の直線を加える
```

分布による比較

以下では正規分布 $N(\mu, \sigma^2)$ の母数 μ の推定を考える．通常この推定には標本平均が用いられるが，μ は母中央値でもあるので標本中央値も推定に用いることが出来る．そこで標本平均と標本中央値を比較することを考える．

問 20．以下のコマンドを実行し，標本平均と標本中央値の期待値やヒストグラムを求めよ．

標本平均と標本中央値の分布．

```
Data <- (NULL)                        # リストの初期化
Mean   <- (NULL)                      # 標本平均のリスト
Median <- (NULL)                      # 標本中央値のリスト
for(i in 1:10000){                    # 以下を 10000 回繰り返す
 Data <- rnorm(20)                    # 正規乱数 20 個を Data に代入
 Mean <- c(Mean,mean(Data))           # Data の標本平均をリストに追加
 Median <- c(Median,median(Data)) # Data の標本中央値をリストに追加
}
```

注意：x 軸の間隔を指定し，重ねがきすると比較しやすい．

平均二乗誤差による比較

問 20 から標本平均と標本中央値の分布のばらつき具合が異なることが分かった．では母平均 0 に平均的に近い値が出ているのはどちらであろうか．その指標の一つが $E((\hat{\theta} - \theta)^2)$ で定義される**平均二乗誤差**である．

問 21．この問題では $\mu = 0$ なので標本平均の平均二乗誤差は `mean(Mean*Mean)` で求めることが出来る．（通常は `Mean` の代わりに `Mean-`θ としたものである．）標本平均と標本中央値の平均二乗誤差を比較し，どちらの推定量のほうがこの意味で優れているか調べよ．

漸近正規性（補足）

漸近正規性とは標本サイズが大きいとき，正規分布による近似が出来ることを表す．正規分布で近似が出来るときは正規分布の%点を用いて推定・検定を行うことができる．

問 22．中心極限定理から標本サイズが大きいとき標本平均の分布は正規分布で近似できる．まず `hist(rchisq(1000,2))` によりカイ二乗分布の母集団分布を確かめよ．次に以下のコマンドで標本平均の分布が正規分布で近似できることを確かめよ．また `N` を $2, 5, 10, 20, 50$ などと変化させ，およそ正規近似が出来そうな最小の `N` を探索せよ．

標本平均の分布．

```
S <- (NULL)                 # 標本平均のリスト
N <- 100                    # 一度の標本抽出における標本サイズ
Data <- (NULL)              # リスト Data の初期化
for(i in 1:1000){           # 以下を 1000 回繰り返す
 Data <- rchisq(N,2)        # 自由度 2 のカイ二乗乱数 N 個を Data に代入
 S <- c(S,mean(Data))       # Data の標本平均をリスト S に追加
}
```

6　区間推定

母平均の 95%信頼区間（母分散既知）

正規母集団 $N(\mu, \sigma^2)$ の独立標本 $X_1, X_2, \ldots, X_n$ の 95%信頼区間は次で求めることができる：

$$\left(\bar{X} - 1.96\sqrt{\frac{\sigma^2}{n}}, \quad \bar{X} + 1.96\sqrt{\frac{\sigma^2}{n}} \right)$$

問 23. 正規分布 $N(0,1)$ に従う乱数を 20 個発生させて，母平均の 95% 信頼区間を求めよ.

母平均の 95%信頼区間.

```
n <- 20                                  # 標本サイズ
Data <- rnorm(n)                         # N(0,1) の乱数 n 個を代入
u <- mean(Data)+1.96/sqrt(n)             # 上側限界を u に代入 (母分散=1 に注意)
l <- mean(Data)-1.96/sqrt(n)             # 下側限界を l に代入
c(l,u)                                   # 母平均の 95%信頼区間を表示
```

信頼係数 95%の意味

さて 95%信頼区間の”95%” とはどういう意味なのだろうか．これは何度も「標本抽出・信頼区間の導出」を繰り返したとき，信頼区間らの 95%は母数を含んでいるという意味である.

問 24. 標本抽出を何度も行うことにより信頼係数を求めてみよう．以下のコマンドを実行し，求めた信頼区間らの何%に推定したい母平均 0 が含まれているかを調べよ.

信頼係数 (被覆確率) を求める.

```
n <- 20; B <- 10000; C <- 0      # 標本サイズ，実験回数，カウンター
for(i in 1:B){                   # 以下を B 回繰り返す
 Data <- rnorm(n)                # N(0,1) の乱数 n 個を代入
 u <- mean(Data)+1.96/sqrt(n) ; l <- mean(Data)-1.96/sqrt(n)
 if(u>0&&l<0) C <- C+1           # もし，信頼区間 (l,u) に 0 が含まれていたら
}                                # カウンターに 1 を加える
C/B                              # 被覆に成功した割合を出力
```

信頼区間のプロット.

```
n <- 50; B <- 100                    # 標本サイズ，実験回数
d <- 5.5/sqrt(n)                     # 信頼区間を描く範囲用
plot(1:B, ylim=c(-d,d),type="n")     # プロット描写用の画面
abline(h=0, col="red")               # 真の母平均 0 を赤で表示
for(i in 1:B){
 Data <- rnorm(Data)
 u <- mean(Data)+1.96/sqrt(n) ; l <- mean(Data)-1.96/sqrt(n)
 segments(i,l,i,u)                   # 信頼区間をバーで表示
 if(u<0||l>0) text(i,d,"x")          # 失敗したら x を表示
}
```

母比率の 95%信頼区間

ここでは二項母集団 $B(n,p)$ の母比率 p の 95%信頼区間を考えよう．ド・モアブル=ラプラスの定理から n が十分大きいとき以下が成り立つ：

$$P\left(\hat{p}-1.96\sqrt{\frac{p(1-p)}{n}} \le p \le \hat{p}+1.96\sqrt{\frac{p(1-p)}{n}}\right)=0.95$$

ここで $\hat{p}$ は n 個の標本のうち" 成功" となった割合である．このままでは，信頼区間の端に未知の p が含まれているので，推定量 $\hat{p}$ で置き換えた次の近似信頼区間がよく用いられる．

$$\left(\hat{p}-1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\quad \hat{p}+1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

問 25．信頼区間の被覆確率を求めよ．また `n <- 10` では近似が悪くなることを確かめよ．

母比率の近似 95%信頼区間.

```
n <- 50                                          # 標本サイズ
p <- 0.4                                         # 推定したい母比率
B <- 10000; C <- 0                               # 実験回数，カウンター
for(i in 1:B){
 Data <- rbinom(n,1,p)                           # B(1,p) の乱数を n 個発生
 u <- mean(Data)+1.96*sd(Data)/sqrt(n)           # 問 26 で書き換える部分
 l <- mean(Data)-1.96*sd(Data)/sqrt(n)           # 問 26 で書き換える部分
 if(u>p&&l<p) C <- C+1                           # 問 26 で書き換える部分
}
C/B                                              # 被覆確率を出力
```

分散安定化変換

前項から標本サイズ n が小さいときは信頼区間の精度が悪いことが分かった．この対処法として様々なものが考えられるが，ここでは次の命題を使った方法を紹介する：

命題：$Y \sim B(n,p)$ であるとき，$X=\sin^{-1}(\sqrt{Y/n})$ と変換すると，n が十分大きいとき $V(X)\approx 1/4n$ となる．(分散安定化の一種で二項変量に対する逆正弦変換と呼ばれる.)

この命題を用いると $\sin^{-1}(\sqrt{p})$ の信頼区間が

$$\left(\sin^{-1}(\sqrt{\hat{p}})-1.96\sqrt{\frac{1}{4n}},\quad \sin^{-1}(\sqrt{\hat{p}})+1.96\sqrt{\frac{1}{4n}}\right)$$

となり，信頼区間の幅が $\hat{p}$ によらず一定となる．p について解けば信頼区間が得られる．

問 26．問 25 の「書き換える部分」を変更し，`n <- 10` 程度でも近似が良いことを確かめよ．

分散安定化変換を用いた信頼区間.

```
u <- asin(sqrt(mean(Data)))+1.96/sqrt(4*n)
l <- asin(sqrt(mean(Data)))-1.96/sqrt(4*n)
if(u>asin(sqrt(p))&&l<asin(sqrt(p))) C <- C+1
```

7　2群の位置母数の比較（その２）

§4 では対応のない 2 群の比較法として t-検定，Wilcoxon の順位和検定について扱った．ここでは対応のある 2 群の比較の検定法である対応のある t-検定，Wilcoxon の符号付順位和検定について扱う．

対応のある 2 群の有意差検定の種類

対応のないときと同様に正規性が仮定できるかどうかで次の 2 種類がある．どちらを用いるかについては，§4 と同様に QQ プロットを用いるが問 27 で述べるように若干注意がいる．

(1) 対応のある t-検定（正規性が仮定できる）

(2) Wilcoxon の符号付順位和検定（正規性が仮定できない）

問 27．§4 ではデータをファイルから読み込んだが今回は直接入力してみよう．データは D.S.Sladsburg(1970) に掲載されていたある精神安定薬を投与する前と投与後の興奮度（Hamilton Scale Factor IV）を測定したものである．これについてどちらの検定を用いたらよいかを判断せよ．

対応のあるデータの正規性の確認.

```
Bef <- c(1.83, 0.50, 1.62, 2.48, 1.68, 1.88, 1.55, 3.06, 1.30)
Aft <- c(0.878, 1.647, 0.598, 2.050, 1.060, 1.290, 1.060, 3.140, 1.290)
Dif <- Aft- Bef              # 投与後から投与前を引いたデータ
qqnorm(Dif)                  # Dif の QQ プロット
```

対応のある t-検定

X_i を投与前，Y_i を投与後のデータとし，$Z_i = Y_i - X_i$ とおく．このとき対応のある t-検定は $Z_i \sim N(\theta, \sigma^2)$ であることを仮定し，次の検定問題

$$H_0 : \theta = 0 \quad \text{v.s.} \quad H_1 : \theta \neq 0$$

に帰着させている．問 27 のデータに正規性が仮定できるのであれば次のコマンドで検定結果が出力される．

対応のある t-検定.

```
t.test(Bef, Aft, paired=T)          # 対応のある 2 群の差の両側検定
```

Wilcoxon の符号付順位和検定（ノンパラメトリック検定）

正規性が仮定できないときは前項の $Z_1, Z_2, \ldots, Z_n$ の中央値が 0 かどうかの検定である Wilcoxon の符号付順位和検定を行う．問 27 のデータでは次のコマンドで検定を行うことが出来る．

Wilcoxon の符号付順位和検定.

```
wilcox.test(Bef, Aft, paired=T)     # 対応のある 2 群の差の両側検定
```

問 28．問 27 のデータに対して適切な方法で検定を行え．また問 2 で作成したデータファイルについて同様の検定を行え．このとき以下の注意を参照すると良い．

ファイル読込における注意

§1 や §4 では Data$SCORE のようにデータ名と列名を併記したがいちいちデータ名を表記するのは面倒で間違いも多い．attach コマンドを用いるとその入力が不要になる．次の方法で §1 や §4 と同様の結果が得られる．

attach の使用例.

```
Data <- read.csv("Data0415.csv")     # Data にデータファイルを代入
attach(Data)                         # 以下では Data について扱う
hist(Score)                          # Data の Score 列のヒストグラム
boxplot(Score[Group==1],Score[Group==0])     # 1 群 0 群のデータの箱髭図
t.test(Score[Group==1],Score[Group==0],var.equal=T) # t 検定
detach(Data)                         # attach 設定を元に戻す
```

片側検定について

§4 および本節では両側検定のみを扱った．片側検定については alternative=''greater'' や alternative=''less'' を加える必要がある．

Student の t-検定（右側検定）.

```
t.test(Case, Cont, alternative="grater")  # 対応のない 2 群の差の右側検定
```

必要な標本サイズ

§3 で見たように標本サイズ n が増加すればするほど検出力が上がり棄却されやすくなる．これは実社会で意味のない差を検出してしまう可能性がある．そのため本来は実験の前に標本サイズを決定しておく必要がある．例えば対応のない 2 群の t-検定においては，有意水準 α，母分散 σ^2，意味のある差 $d = \mu_B - \mu_A$ とするとき，検出力を $1-\beta$ 以上にするには

$$n \geq \frac{2\sigma^2(z_{\alpha/2}+z_\beta)^2}{d^2}, \qquad \left(\int_{z_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = \alpha\right)$$

となる必要がある．なお，σ^2 は未知なので見積もっておく必要がある．

問 29．ある調査班は，都市部の A 地区と農村部の B 地区で住民の総コレステロール値の差があるかどうかを調査することになった．両群の差が 10 以上あれば地域により差があると判断できるとする．標準偏差は 34.0 程度と考えられる．有意水準 5% の検定を行うとして，検出力を 90% 確保するには，標本数を各群いくつずつにすればよいかを求めよ．R では qnorm で標準正規分布の累積確率が求まるので上側 0.025 点は qnorm(1-0.025) とすればよい．

標本サイズのための関数の定義.

```
samplesize <- function(a,b,d,s){  # a,b,d,s が引数の関数 samplesize を定義
return(2*s^2*(qnorm(1-a/2)+qnorm(b))^2/(d^2))
}                                 # samplesize は引数からこの値を返す
samplesize(0.05,0.90,10,34.0)     # 必要な標本サイズを出力
```

注意．power.t.test(power=0.9,delta=10,sd=34) としてもよい．

8　多群の位置母数の比較

§4 および §7 では 2 群の比較について扱ったが，実際の問題として 3 つ以上の群を比較することもある．ここではそのような場合に用いられる一元配置分散分析と多重比較について簡単に扱う．

一元配置データ

次は日照条件などが等しいいくつかの畑でそれぞれ肥料 A_1, A_2, A_3 を用いたときの収穫高のデータである．このことから肥料によって有意な差があるといえるだろうか．

表 2．収穫高のデータ

肥料						平均	SD
A_1	2.9	3.2	3.2	2.6	3.0	2.98	0.25
A_2	2.8	2.5	2.7	2.6	2.9	2.70	0.16
A_3	3.1	2.8	3.2	3.3		3.10	0.22

問 30．表 2 のデータを入力し，図によって比較してみよう．このとき水準 A_1, A_2, A_3 のいずれのデータなのかが分かるようなグループ変数を入れる必要があることに注意せよ．また参考書のデータを Excel または直接入力により R に読み込め．

一元配置データの入力.

```
Group <- c(1,1,1,1,1,2,2,2,2,2,3,3,3,3)      # グループ変数
Data <- c(2.9, 3.2, 3.2, ..., 3.2, 3.3)      # 収穫高（途中を省略している）
plot(Group,Data)                             # 群ごとの散布図
```

一元配置分散分析

等分散である 3 つの正規母集団 $N(\mu_1, \sigma^2), N(\mu_2, \sigma^2), N(\mu_3, \sigma^2)$ の平均に関する検定問題

$$H_0 : \mu_1 = \mu_2 = \mu_3 \quad \text{v.s.} \quad H_1 : \text{not} H_0$$

を考えよう．ここで母集団 i からの n_i 個の標本を $y_{i1}, y_{i2}, \ldots, y_{in_i} \overset{i.i.d.}{\sim} N(\mu_i, \sigma^2)\,(i = 1, 2, 3)$ とする．さらに全データの平均を $\bar{y}$，それぞれの群の標本平均を $\bar{y}_i$ とおく．このとき次の平方和分解ができる：

$$\sum_{i=1}^{3}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{3}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{3} n_i(\bar{y}_i - \bar{y})^2$$

ここで左辺は**全平方和** S，右辺第 1 項を**郡内平方和** W，右辺第 2 項を**群間平方和** B という．さて H_0 が正しいとき，$\bar{y}_i$ は $\bar{y}$ と近いはずだから，W と S は近いはずである．よって

$$F = \frac{B/(3-1)}{W/(n-3)} > F_{3-1,n-3}(0.05) \quad \Rightarrow \quad H_0\text{を棄却}$$

という検定方式が考えられる．これを**一元配置分散分析**という．

問 31．R では `oneway.test` や `aov` で一元配置分散分析を行うことが出来る．問 30 のデータに対して一元配置分散分析を行い，肥料によって生産高に差があるかどうかの判断を下せ．

一元配置分散分析.

```
oneway.test(Data~Group, var=T)     # 引数はデータ名~グループ変数
```

Kruskal-Wallis 検定

前項の一元配置分散分析は §4 の Student の t 検定と同様に正規性と等分散という 2 条件が必要である．これらの仮定が満たされないときは 3 群の中央値が等しいかどうかの検定である **Kruskal-Wallis 検定**を行う．

問 32．それぞれの正規母集団 $N(\mu_i, \sigma_i^2)$ の母分散に関する検定問題

$$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 \quad \text{v.s.} \quad H_1 : \text{not} H_0$$

検定法として **Bartlett 検定**がある．問 30 のデータに対して Bartlett 検定を行い等分散正が仮定できるか判断せよ．

Bartlett 検定 (3 つ以上の母分散の検定).

```
bartlett.test(Data, Group)         # 引数はデータ名, 群変数
```

問 33．等分散性が満たされないときの Kruskal-Wallis 検定は `kruskal.test` で行うことが出来る．問 32 の結果に関わらず，問 30 のデータに対して Kruskal-Wallis 検定を行え．

Kruskal-Wallis 検定.

```
kruskal.test(Data~Group)           # 引数はデータ名~群変数
```

多重比較

一元配置分散分析や Kruskal-Wallis 検定は全体で差があるかどうかを見るものである．一方各群の組み合わせで平均値を比較する**多重比較法**も多くの種類がある．

問 34．分散分析と同様に正規性と等分散性が仮定できるときは Tukey の HSD 法がある．また Bonferroni の方法の改良版である Holm の多重比較もよく行われる．問 30 のデータに対して多重比較を行い，差のあるペアがあれば検出せよ．

Tukey の HSD 法.

```
Group2 <- factor(Group2)           # HSD を使うためには因子指定の必要あり
TukeyHSD(aov(Data~Group2))         # 引数は aov(データ名~因子)
```

Holm の多重比較法.

```
pairwise.t.test(Data, Group, p.adj="holm")      # p.adj で多重比較法を指定
```

問 35．問 34 のノンパラ版として `pairwise.wilcox.test` がある．問 30 のデータに対して多重比較を行い，差のあるペアがあれば検出せよ．

ノンパラ版多重比較.

```
pairwise.wilcox.test(Data, Group, p.adj="holm") # p.adj で多重比較法を指定
```