統計学・データ分析勉強会／第 1 回宿題解答例

第 1 回宿題の解答例を以下に示します。いずれの問題も手計算で答えを求めることが可能ですが、解答例は R を使って作成しました。R の基本操作の学習に役立てて下さい。一部、初学者には高度な部分がありますので、全てを理解できなくても構いません。

最初に以下のコードをコンソールに入力して下さい。R は通常、桁数の大きい数字を特別な数学表記で出力します（例えば 200000 は 2e+05 と表示します）が、今回は特別な命令によってその機能を停止します。

```
options(scipen=100)  # 数字の表記方法を変えるための命令
```

問 1

まず、平均月収と実質手取りのデータをそれぞれ Gross_Income と Net_Income という名前の変数（ベクトル）に格納します。

```
Gross_Income <- c(314600, 403500, 348400, 360800, 277300, 353700, 190200, 478200,
            465700, 339500, 80000, 39100, 302200, 80000, 74600, 44400, 163500,
            129700, 32000, 179500, 110200, 35500, 46200, 28400)
Net_Income <- c(240000, 280000, 270000, 270000, 190000, 240000, 140000, 270000,
            340000, 280000, 55000, 35000, 220000, 75000, 60000, 40000, 160000,
            110000, 30000, 140000, 95000, 30000, 40000, 25000)
```

length() という関数を使い、作成したデータのサイズを確認しておきましょう。どちらも 24 個のデータが入っています。

```
length(Gross_Income)
```

```
## [1] 24
```

```
length(Net_Income)
```

```
## [1] 24
```

残念ながら、R には度数分布表を作る関数が用意されていません。このため、ここでは自分で関数を定義することにします。以下のコードをコンソールに入力して下さい（各行の「#」以降は説明のためのコメントでコードの実行内容には影響しないため、無視して OK です）。現時点では詳しい中身を理解する必要はありません。なお、下記コードは http://rplus.wb-nahce.info/rsemi_stat_basic/r_dosuubunnpuhyou_sakusei.html を修正のうえ転用しました。

```
freqtab <- function( # 度数分布表を作成する関数を定義する
        x, # 度数分布表のもとになるデータ
        cn=1+log2(length(x)), # 階級の数は「スタージェスの公式」で決める
        wid=FALSE ) # 階級の幅
{
    x <- x[!is.na(x)] # 欠損値を除く
    min <- min(x) # データの最小値
    max <- max(x) # データの最大値
    ran <- max - min # データのレンジ（範囲）
    if(!wid){ wid <- round(abs(ran/cn), 3) } # 階級の幅を指定しない場合は自動で計算
    cla <- floor(x/wid) # データを階級に振り分ける
    mnc <- min(cla) # 階級の最小値
    mxc <- max(cla) # 階級の最大値
    cla <- factor(cla, levels=mnc:mxc) # データを階級ごとに集計
    freq <- table(cla) # 階級ごとの度数を数える
    names(freq) <- paste(mnc:mxc*wid, "-", mnc:mxc*wid+wid) # 階級の名称を定義する
    percent <- round(freq/sum(freq)*100,2) # 相対度数（%）
    cum.pcnt <- round(cumsum(percent),2) # 累積相対度数（%）
    return(cbind(freq, percent, cum.pcnt)) # 度数分布表を返す
}
```

これで、度数分布表を作るのに便利な `freqtab()` という関数ができました。

それでは早速、平均月収の度数分布表を作成しましょう。階級の幅は 100000 円に指定します。

```
freqtab(Gross_Income, wid=100000) # 度数分布表を作成。階級の幅は 100000 に指定
```

```
##                  freq percent cum.pcnt
## 0 - 100000          9   37.50    37.50
## 100000 - 200000     5   20.83    58.33
## 200000 - 300000     1    4.17    62.50
## 300000 - 400000     6   25.00    87.50
## 400000 - 500000     3   12.50   100.00
```

上の度数分布表において、`freq` が度数、`percent` が相対度数（%）、`cum.pcnt` が累積相対度数（%）を表しています。

実質手取りの度数分布表も同じように作成しましょう。

```
freqtab(Net_Income, wid=100000) # 度数分布表を作成。階級の幅は 100000 に指定
```

```
##                  freq percent cum.pcnt
## 0 - 100000         10   41.67    41.67
## 100000 - 200000     5   20.83    62.50
## 200000 - 300000     8   33.33    95.83
## 300000 - 400000     1    4.17   100.00
```

一方、ヒストグラムは `hist()` という関数で簡単に作成できます。単純に `hist(Gross_Income)` とすれば、平均月収のヒストグラムができます。

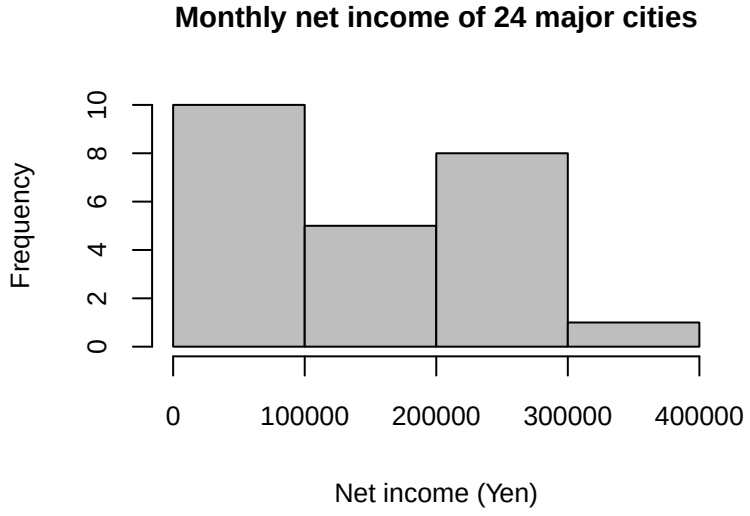
```
hist(Gross_Income)  # ヒストグラムを描く
```

`hist` 関数に渡す変数を変えることで、ヒストグラムの階級の数やラベルなどを指定できます。平均月収のヒストグラムを見やすくするため、タイトルとラベル、色を変更してみましょう。

```
# タイトル、横軸のラベル、縦軸のラベル、ヒストグラムの色、縦軸の範囲、文字サイズを指定する
hist(Gross_Income, main="Monthly gross income of 24 major cities",
     xlab="Gross income (Yen)", ylab="Frequency", col="grey", ylim=c(0, 10),
     cex.main=0.9, cex.axis=0.8, cex.lab=0.8)
```

同様に実質手取りのヒストグラムも作成しましょう。ここでは階級の数も指定しています。

```
# 階級の数、タイトル、横軸のラベル、縦軸のラベル、ヒストグラムの色、縦軸の範囲、文字サイズを指定する
hist(Net_Income, breaks=4, main="Monthly net income of 24 major cities",
     xlab="Net income (Yen)", ylab="Frequency", col="grey", ylim=c(0, 10),
     cex.main=0.9, cex.axis=0.8, cex.lab=0.8)
```



問 2

A 国の月収調査のデータと B 国の月収調査のデータを、それぞれ IncA、IncB という名前の変数に格納します。

```
IncA <- c(530000, 520000, 510000, 500000, 490000, 480000, 470000)
IncB <- c(1200000, 900000, 800000, 500000, 60000, 30000, 10000)
```

平均値は関数 mean()、標準偏差は関数 sd() を使って求めることができます。

A 国の月収の平均値と標準偏差を計算します。

```
mean(IncA)
```

```
## [1] 500000
```

```r
sd(IncA)
```

```
## [1] 21602.47
```

B 国の月収の平均値と標準偏差を計算します。

```r
mean(IncB)
```

```
## [1] 500000
```

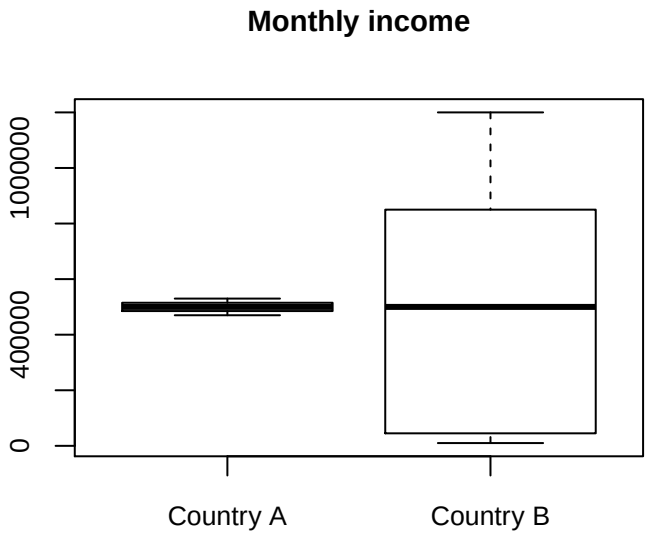
```r
sd(IncB)
```

```
## [1] 482113.4
```

以上より、**A 国と B 国の月収の平均値は同じですが、B 国の方が圧倒的にばらつき（標準偏差）が大きい**ことがわかります。

ちなみに、この結果は `boxplot()` という関数を使って「箱ひげ図」で表示すると見た目にも分かりやすくなります。以下のように、A 国は B 国に比べて狭い月収の範囲にデータが集中していることが明らかです。

```r
boxplot(IncA, IncB, # 箱ひげ図を描く
        names=c("Country A", "Country B"), main="Monthly income",  # ラベルやタイトルを指定
        cex.main=0.9, cex.axis=0.8, cex.lab=0.8) # 文字のサイズを指定
```



なお、`summary()` という関数を使うとデータの平均値、中央値、最小値、最大値、四分位点をまとめて表示できます。

```r
summary(IncA)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  470000  485000  500000  500000  515000  530000
```

```r
summary(IncB)
```

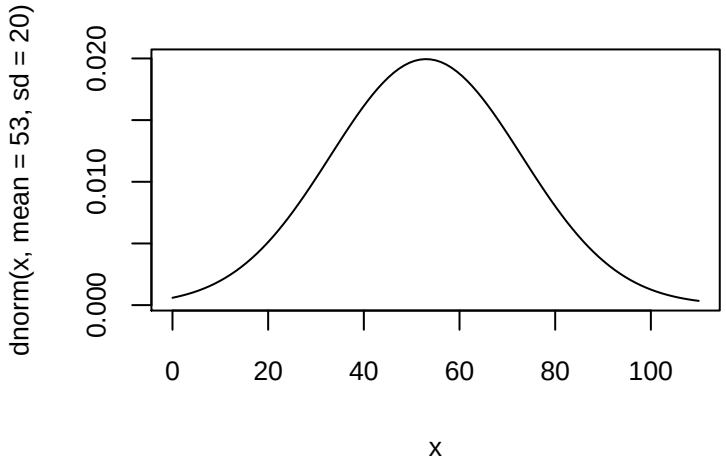
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10000   45000  500000  500000  850000 1200000
```

問 3

正規分布（Normal Distribution）は `dnorm()` という関数に平均値 `mean` と標準偏差 `sd` を渡すことで作成できます。また、ある関数をグラフで表示したいときは、`curve()` という関数を使います。これらを組み合わせることで正規分布のグラフを描くことができます。

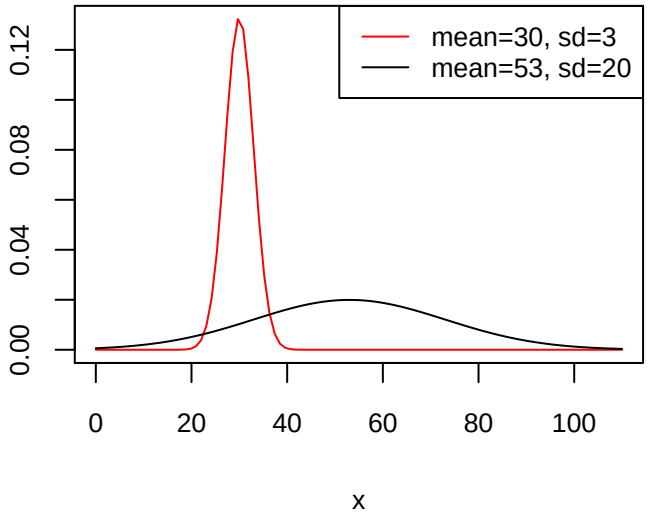
平均値 53、分散 400 の正規分布のグラフは以下のようになります。標準偏差は $\sqrt{400} = 20$ であることに注意しましょう。

```
curve(dnorm(x, mean=53, sd=20), # 平均値 53、分散 400 の正規分布を描く
      from=0, to=110, # 横軸は 0 ～ 110 に指定
      cex.axis=0.8, cex.lab=0.8)  # 文字のサイズを指定
```



次に平均値 30、分散 9（標準偏差は $\sqrt{9} = 3$）の正規分布のグラフを描きます。なお、ここでは上で描いた正規分布との違いが分かりやすいように、2 つのグラフを重ねて表示することにします。グラフの縦軸が先ほどと変わっている点に注意して下さい。

```
curve(dnorm(x, mean=30, sd=3), # 平均値 30、分散 9 の正規分布を描く
      from=0, to=110, col="red", ylab="", # グラフの範囲やラベルを指定。色は赤
      cex.main=0.9, cex.axis=0.8, cex.lab=0.8) #  文字サイズを指定
curve(dnorm(x, mean=53, sd=20), add=TRUE) #  平均値 53、分散 400 の正規分布を重ねて表示
label <- c("mean=30, sd=3", "mean=53, sd=20")  # 凡例に表示するラベルを定義
color <- c("red", "black") # 2 つのグラフの色分け
legend("topright", legend=label, col=color, lty=c(1, 1), cex=0.8)  # 凡例を作成
```



上のグラフから、**正規分布は分散が小さいほどグラフの幅が狭まり、背が高くなる**ことが分かります。なお、正規分布は確率密度を表しているため、どちらのグラフも面積（曲線と x 軸で囲まれた範囲の面積）を合計すると 1 になります。

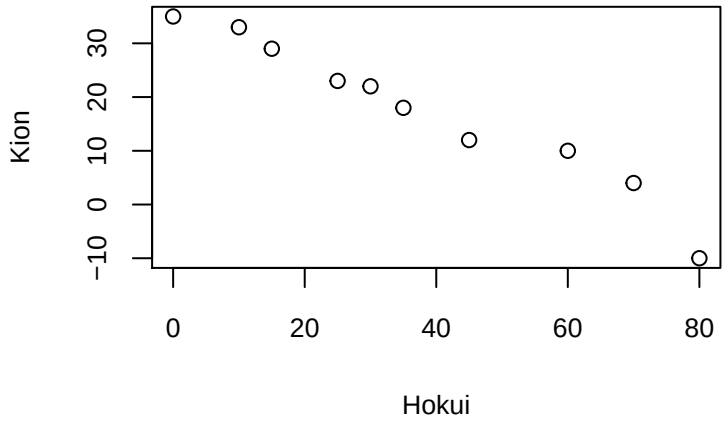
問 4

北緯と平均気温をそれぞれ `Hokui`、`Kion` という名前の変数として定義します。

```
Hokui <- c(0, 35, 30, 70, 10, 45, 60, 80, 15,  25)
Kion <- c(35, 18, 22, 4, 33, 12, 10, -10, 29, 23)
```

散布図は `plot()` 関数を使って描くことができます。

```
plot(Hokui, Kion, cex.axis=0.8, cex.lab=0.8)
```



上の散布図から、**「北緯が高いほど平均気温が低下する」という明確な傾向**が分かります。

相関係数（Correlation Coefficient）は `cor()` という関数で計算します。

```
cor(Hokui, Kion)
```

```
## [1] -0.9804038
```

なお、相関係数について少しだけ丁寧に触れておきましょう。「北緯」と「平均気温」のような n 個のデータの組が (x_1, y_1), (x_2, y_2), $\cdots$, (x_n, y_n) と与えられた場合、変数 x と y の間の相関係数 r_{xy} は以下の式で計算します。

$$r_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2}\sqrt{\Sigma(y_i - \bar{y})^2}}$$

ここで、$\bar{x}$ と $\bar{y}$ はそれぞれ x と y の平均値を表します。r_{xy} の分子に注目すると、r_{xy} は「x が増大すれば y が増大する」傾向にあるデータではプラスになり、逆に「x が増大すれば y が減少する」傾向にあるデータではマイナスになることが予想できます。相関係数は常に $-1 \leq r_{xy} \leq 1$ の値をとることが知られており（証明はやや難しいので、興味のある人は統計の教科書やウェブサイトを参照してください）、$r_{xy} = 1$ に近いほど「正の相関が強い」、$r_{xy} = -1$ に近いほど「負の相関が強い」と表現します。

上で計算したように北緯と平均気温の相関係数は -0.98 ですから、2 つの変数は**非常に強い負の相関をもつ**と言えます。

問 5

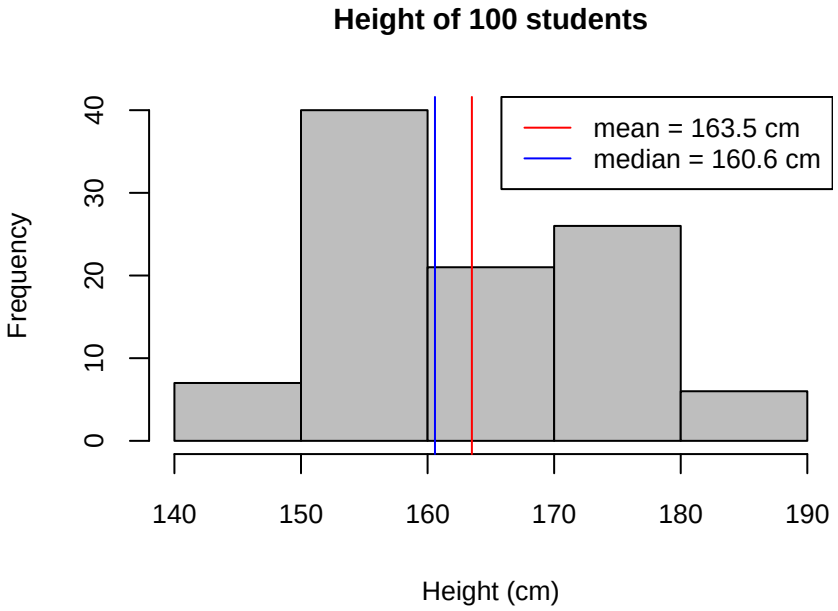
(1) **誤り（正しいとは限らない）**
(2) **正しい**
(3) **誤り（正しいとは限らない）**

【理由】

この問題には複数の答え方があるので、論理的に理由を説明できていれば正解とします。

一般的に言って、正規分布のような理想的な分布の場合を除き、データの平均値と中央値（真ん中の値）は必ずしも一致しません。従って、身長が平均値（163.5cm）よりも高い生徒の数と低い生徒の数が 50 人ずつ同数になるという仮説 (1) は誤りです。また、分布が大きく歪んでいるデータをヒストグラムや度数分布表で表示すると、平均値を含む階級の度数が他の階級の度数より小さくなるケースもあります。このため仮説 (3) も誤りです。一方、仮説 (2) は平均値の定義そのものですから（小数点の四捨五入の影響を無視すれば）常に正しいと言えます。

仮説 (1) と (3) が誤りである理由を、仮想的に作成したデータを使って確認しましょう。例えば、生徒 100 人の身長が、以下のヒストグラムのような分布になっていると仮定します。この分布では 100 人の身長の平均値は 163.5cm、中央値は 160.6cm です。平均値が中央値より大きいため、身長が平均値（163.5cm）よりも高い生徒の数は、低い生徒の数よりも少なくなります。このように、仮説 (1) は必ずしも成り立ちません。また、ヒストグラムを見ると、「身長が 160cm 以上で 170cm 未満の生徒」の数は、「150cm 以上で 160cm 未満の生徒」や「170cm 以上で 180cm 未満の生徒」の数よりも少ないことが明らかです。従って仮説 (3) も誤りとなります。



ただ、多数の人の身長データは一般的にほぼ正規分布に従うと言われています。今回の問題でも、もし 100 人の身長データが正規分布のように中央部分に集中した左右対称の分布をしているという前提に立てば、仮説 (1) や (3) が成り立つ可能性も十分にあります。